FACULTAT D'INFORMÀTICA DE BARCELONA

BACHELOR'S DEGREE IN INFORMATICS ENGINEERING

Speed Dating

DATA MINING COURSE

Members:

Pol Pérez Castillo

Maxime Côté-Lapointe

Daniel Muñoz Arroyo

Alejandro Salvat Navarro

Index

Description of the problem

This dataset comes from an experiment based on speed dating events from 2002 to 2004. During the events, participants had a four-minute "first date" with each person of the opposite sex. At the end of the four minutes, they were asked if they would like to see their date again. This dataset includes information about the participants, ratings they have made and received on each of their dates, and whether or not they ultimately decided to see each date again.

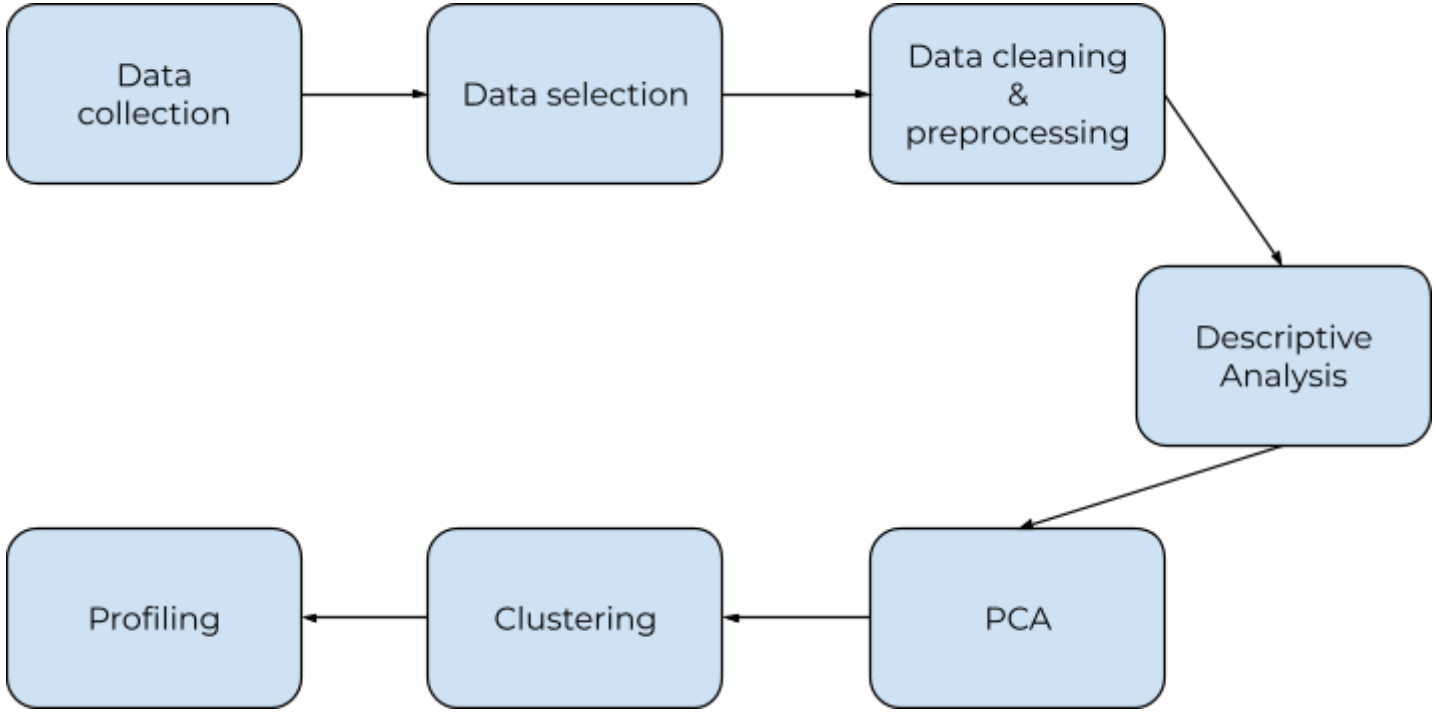
Data source presentation

The dataset has been selected from *Kaggle*, a site where a large community of data scientists from all over the world share different problems, propose their own solutions and discuss different ways and techniques to address the obstacles they face. The file format is present on the website in .csv format, so once downloaded, it is perfectly usable without any additional steps.

In our problem, we have a dataset with 8378 individuals and 123 attributes for each of them, of which we will keep only the 25 apparently most representative attributes in order to reduce the computation time of our algorithms and thus make the academic exercise more comfortable.

Data Mining Process Performed

The data mining process of our project starts with the selection of the dataset we will work with from the Kaggle page. Once the dataset is selected, we proceed to visualise the data included in it, by which we decided to make a selection of 25 variables (the apparently most significant ones) so that we can reduce the complexity of the academic exercise. Once the variables have been selected, we proceed to clean the data and preprocess them, changing the data format to one that is more comfortable to work with and using two different techniques for imputation of missing values: Median imputation and MCA. Once we have completed the tasks related to the pre-processing of our dataset, we go on to perform a Descriptive Analysis of the variables, in addition to performing the PCA process, so that we can see the separability of our data. Finally, we went on to perform the herarchial clustering process of our individuals to see how they could be grouped and the profiling of our variables, so that we could analyse them individually. This whole process is reflected in the following workflow:



Metadata

Variable	modalities	meaning	Type	Measuring unit	Missing code	Range	Role
gender		**Gender at birth of self**	Binary		NA		Explanatory
	M	Male					
	F	Female					
d_age	-	**Age difference between the partners**	Numerical	Years	NA		Explanatory
samerace		**Do the partners have the same race?**	Binary		NA		Explanatory
	0	No					
	1	Yes					
pref_o_attractive		**How important does the partner rate attractiveness?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					
	2	[10,20)					
	3	[20,30)					
	4	[30,40)					
	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
pref_o_sincere		**How important does the partner rate sincerity?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					
	2	[10,20)					
	3	[20,30)					

	4	[30,40)					
	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
pref_o_intelligence		**How important does the partner rate intelligence?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					
	2	[10,20)					
	3	[20,30)					
	4	[30,40)					
	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
attractive_o	-	**Rating by partner (about me) at night of event on attractiveness**	Numerical		NA	[0,10]	Explanatory
sinsere_o	-	**Rating by partner (about me) at night of event on sincerity**	Numerical		NA	[0,10]	Explanatory
intelligence_o	-	**Rating by partner (about me) at night of event on intelligence**	Numerical		NA	[0,10]	Explanatory
attractive_important		**How much importance do you attribute to attractiveness?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					

	2	[10,20)					
	3	[20,30)					
	4	[30,40)					
	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
sincere_important		**How much importance do you attribute to sincerity?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					
	2	[10,20)					
	3	[20,30)					
	4	[30,40)					
	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
intelligence_important		**How much importance do you attribute to intelligence?**	Categorical		NA	[0,100]	Explanatory
	1	[0,10)					
	2	[10,20)					
	3	[20,30)					
	4	[30,40)					

	5	[40,50)					
	6	[50,60)					
	7	[60,70)					
	8	[70,80)					
	9	[80,90)					
	10	[90,100]					
attractive	-	**Rate your attractiveness**	Numerical		NA	[0,10]	Explanatory
sincere	-	**Rate your sincerity**	Numerical		NA	[0,10]	Explanatory
intelligence	-	**Rate your intelligence**	Numerical		NA	[0,10]	Explanatory
attractive_partner	-	**Rate your partner's attractiveness**	Numerical		NA	[0,10]	Explanatory
intelligence_partner	-	**Rate your partner's sincerity**	Numerical		NA	[0,10]	Explanatory
sincere_partner	-	**Rate your partner's intelligence**	Numerical		NA	[0,10]	Explanatory
interests_correlate	-	**Correlation between a participant's and partner's rating of interests.**	Numerical		NA	[-1,1]	Explanatory
expected_num_interested_in_me	-	**Out of the 20 people you will meet, how many do you expect will be interested in dating**	Numerical		NA	[0,20]	Explanatory
like	-	**Did you like your partner? Rating**	Numerical		NA	[0,10]	Explanatory
guess_prob_liked	-	**How likely do you think it is that your partner likes you?**	Numerical		NA	[0,10]	Explanatory
decision		**Decision at night of event.**	Binary		NA		Explanatory
	0	No					
	1	Yes					
decision_o		**Decision of partner at night of event.**	Binary		NA		Explanatory
	0	No					
	1	Yes					
match		**Does the person have a match?**	Binary		NA		Response
	0	No					

	1	Yes					

Preprocessing

In this section it is explained the decisions that we have made in the preprocessing of the data frame.

First step

Our dataset consists of 123 columns and in order to do the work with all of that, it will be so large. As explained in the laboratory sessions, we have made groups of similar characteristics and erased some random columns to reduce from 123 columns to 25.

The groups are the next:

- **Age**: there are 4 variables talking about the age of the person, the age of the partner's person and 1 categorical and 1 numerical with the difference of ages, so we only will keep the difference between ages because it could have more information than only with the age of one person. We will choose the linear because it follows a normal distribution looking on kaggle description.

- **Preference of the partner**: importance of the partner about the other person be attractive, sincere, intelligent, funny, ambitious and shared interests.

- **Rating of partner**: rating of the partner about the other person be attractive, sincere, intelligent, funny, ambitious and shared interest,

- **Importance**: importance about a person being attractive, sincere, intelligent, funny, ambitious and shared interests

- **Self rating**: rating yourself about attractive, sincere, intelligent, funny, ambitious

- **Rating partner**: rating the partner about attractive, sincere, intelligent, funny, ambitious.

- **Self interests about activities**: rating about self activities interest. The activities are the following: sports, tvsports, exercise, dining, museums, art, hiking, gaming, clubbing, reading, tv, theater, movies, concerts, music, shopping, yoga.

- **Things expected**: expected results about if I will be happy, number interested in me and number of matches.

- **Likes**: Rating about if you like the partner and expected like of your partner about you.

- **Others**: Have you met the partner before?, decision at night of event and decision of the partner at night of event, match.

- **Other characteristics**: wave, gender, race of self and race of the partner, samerace, importance same race of self and the partner, importance same religion, field of study.

We have 11 groups and we have chosen 2 or 3 of different variables in each group:

- **Age**: numerical age difference.

- **Preference of the partner**: In kaggle the preference distribution is more or less the same in attractive, sincere, intelligence and funny. We have chosen 3 random. Those are attractive, sincere and intelligence (categorical variables to the importance and preference of each one and the other numerical).

- **Self Interests activity**: we think that we have this information in only the interest correlated, so we haven't used any of them.

- **Things expected**: we have chosen expected number interested in me.

- **Likes**: like and expected like. We keep both.

- **Others**: we have deleted the met variable because 351 of 7995 answers yes and we keep decision and decision_o.

- **Other characteristics**: we keep same race and gender.

Now we talk about what we have done with each variable to make it usable.

gender

There were useless characters in each cell, so we've done a refactor of all the values. We replaced the "b'male'" values into a "M" and the "b'female'" values into a "F". No more preprocessing was done because there were no missing values.

d_age

No preprocessing was needed because there were no missing values.

samerace

We replaced the "b'0'" values into a "0" and the "b'1'" values into a "1". No more preprocessing was done because there were no missing values.

pref_o_attractive

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method.

pref_o_sincere

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method

pref_o_intelligence

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method

attractive_o

We have imputed the values in the individuals with NA with the mean method.

sincere_o

We have imputed the values in the individuals with NA with the mean method.

intelligence_o

We have imputed the values in the individuals with NA with the mean method.

attractive_important

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method

sincere_important

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method

intelligence_important

It was a numerical variable and we have factored it and changed its values to ones more in line with the rest of the variables. We have imputed the values in the individuals with NA with the MCA method

attractive

We have imputed the values in the individuals with NA with the mean method.

sincere

We have imputed the values in the individuals with NA with the mean method.

intelligence

We have imputed the values in the individuals with NA with the mean method.
attractive_partner

We have imputed the values in the individuals with NA with the mean method.

intelligence_partner

We have imputed the values in the individuals with NA with the mean method.

sincere_partner

We have imputed the values in the individuals with NA with the mean method.

interests_correlate

We have imputed the values in the individuals with NA with the mean method.

expected_num_interested_in_me

We have deleted this variable because it had 79% of null values.

like

We have imputed the values in the individuals with NA with the mean method.

guess_prob_like

We have imputed the values in the individuals with NA with the mean method.

decision

We replaced the “b’0’” values into a “0” and the “b’1’” values into a “1”. No more preprocessing was done because there were no missing values.

decision_o

We replaced the “b’0’” values into a “0” and the “b’1’” values into a “1”. No more preprocessing was done because there were no missing values.

match

We replaced the “b’0’” values into a “0” and the “b’1’” values into a “1”. No more preprocessing was done because there were no missing values.

Data errors

Now we have a look at the data, to see if there are some errors:
For the numerical data we did a summary of the dataframe to see the ranges of the data as shown in the picture:

```
> summary(df[numericalvariables])
     d_age         attractive_o      sinsere_o      intelligence_o     attractive       sincere         intelligence    attractive_partner intelligence_partner
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 2.000   Min.   : 2.000   Min.   : 2.000   Min.   : 0.000     Min.   : 0.000
 1st Qu.: 1.000   1st Qu.: 5.000   1st Qu.: 6.000   1st Qu.: 7.000   1st Qu.: 6.000   1st Qu.: 8.000   1st Qu.: 7.000   1st Qu.: 5.000     1st Qu.: 7.000
 Median : 3.000   Median : 6.000   Median : 7.000   Median : 7.000   Median : 7.000   Median : 8.000   Median : 8.000   Median : 6.000     Median : 7.000
 Mean   : 4.186   Mean   : 6.186   Mean   : 7.169   Mean   : 7.356   Mean   : 7.084   Mean   : 8.291   Mean   : 7.696   Mean   : 6.185     Mean   : 7.356
 3rd Qu.: 5.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 8.000     3rd Qu.: 8.000
 Max.   :37.000   Max.   :10.500   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000     Max.   :10.000
 sincere_partner  interests_correlate      like       guess_prob_liked
 Min.   : 0.000   Min.   :-0.8300     Min.   : 0.00   Min.   : 0.0
 1st Qu.: 6.000   1st Qu.:-0.0100     1st Qu.: 5.00   1st Qu.: 4.0
 Median : 7.000   Median : 0.2000     Median : 6.00   Median : 5.0
 Mean   : 7.169   Mean   : 0.1923     Mean   : 6.13   Mean   : 5.2
 3rd Qu.: 8.000   3rd Qu.: 0.4300     3rd Qu.: 7.00   3rd Qu.: 7.0
 Max.   :10.000   Max.   : 0.9100     Max.   :10.00   Max.   :10.0
```

The variable attractive_o is a rating betwen 0 and 10 about the atractiveness of his/her partner. As we can see, the maximum is 10.5. This is an error and we will assume that it should be the maximum value (10).

As we can see, the rest of the variables are the spected values because they are in the respective ranges shown in the metadata file.

For the categorical variables, we have those values:

	pref_o_attractive
1	4
2	7
3	2
6	6
10	10
102	5
104	3
108	1
1665	9
5037	8

	attractive_important
1	2
11	5
21	4
31	3
71	1
111	7
151	6
191	10
1827	9
5625	8

	pref_o_sincere
1	3
2	1
3	2
204	4
1661	5
4935	7

	sincere_important
1	3
11	1
21	2
562	4
1787	5
4845	7

	pref_o_intelligence
1	3
2	1
3	2
6	4
507	5
516	6

	intellicence_important
1	3
21	4
91	2
111	1
233	5
377	6

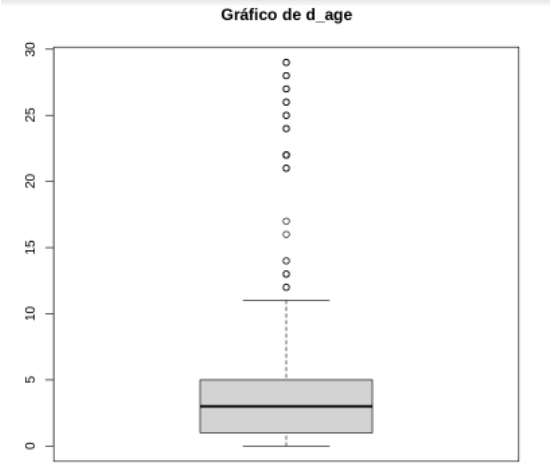
As we can see, the labels are the expected ([0,10]). Those ranges are shown in the metadata file.

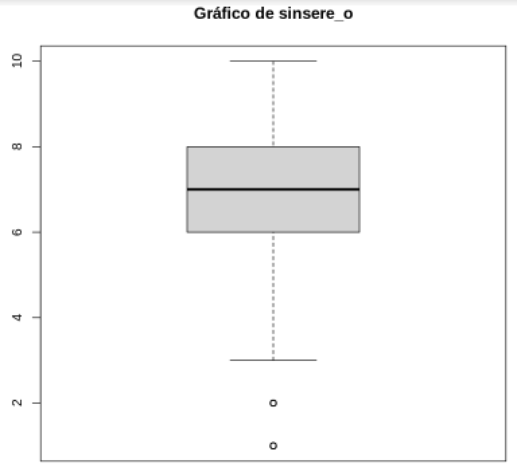
Finally, we look on the binary variables for errors:

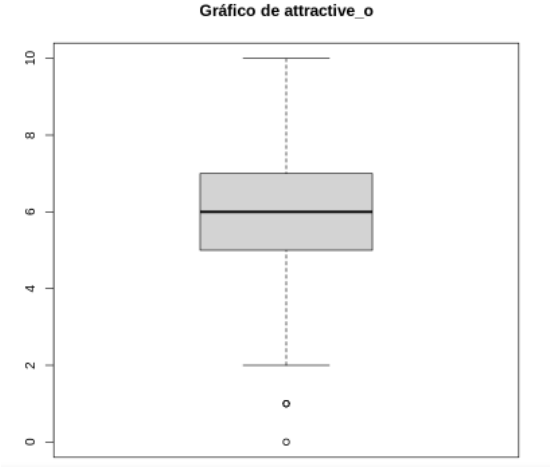
```
     gender
1         F
101       M
  samerace
1         0
3         1
  decision
1         1
6         0
  decision_o
1          0
3          1
  match
1     0
3     1
```

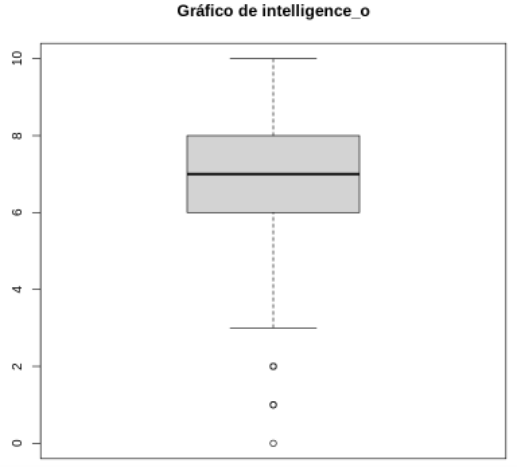
As we can see, in the gender there are only two possible values M or F and in the other variables only can be 0 or 1. Those are expected values according to the metadata file

Outliers









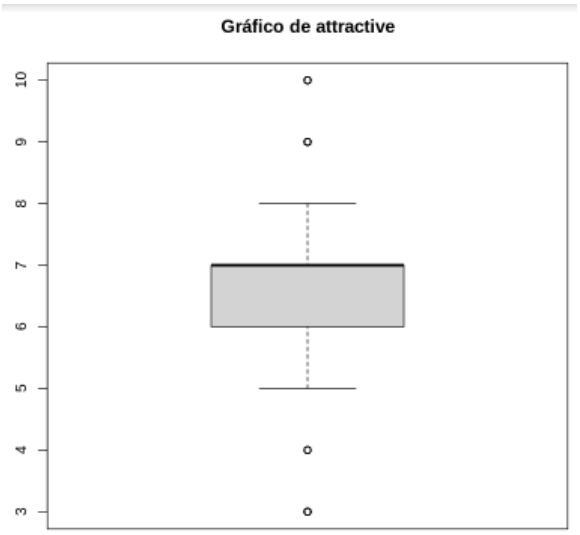


Gráfico de attractive
10
9
8
7
6
5
4
3

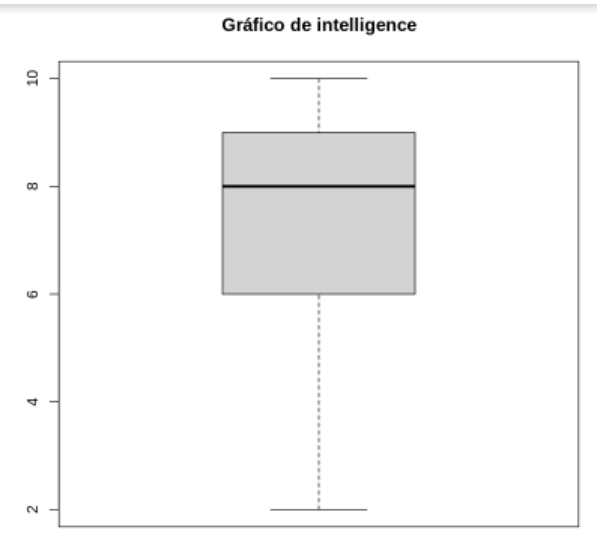


Gráfico de intelligence
10
8
6
4
2

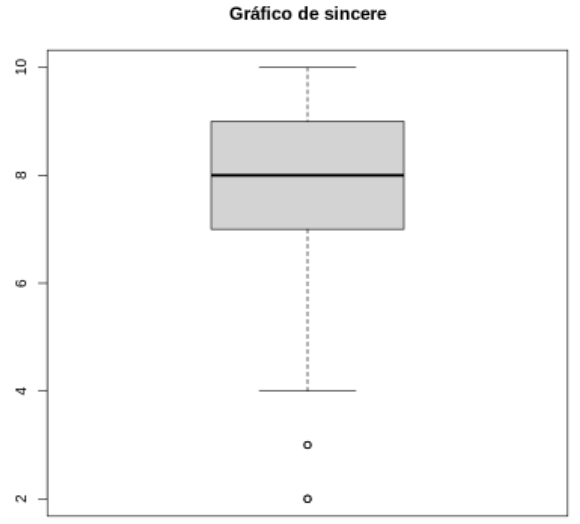


Gráfico de sincere
10
8
6
4
2

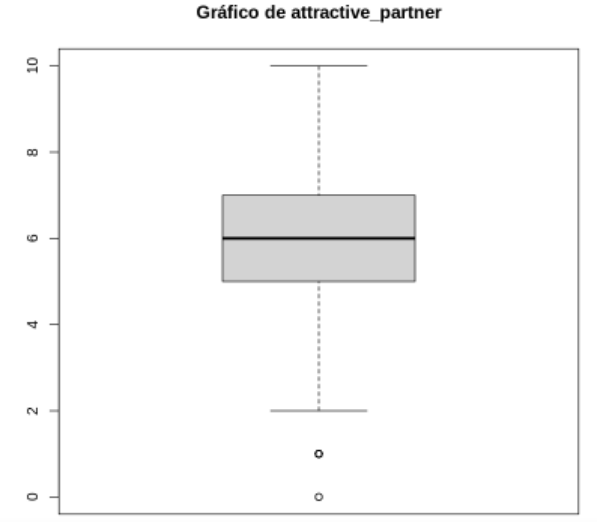
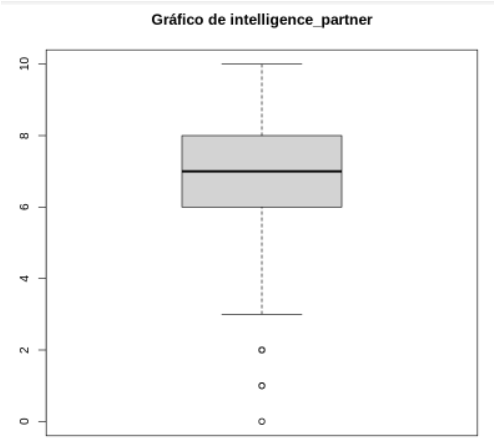
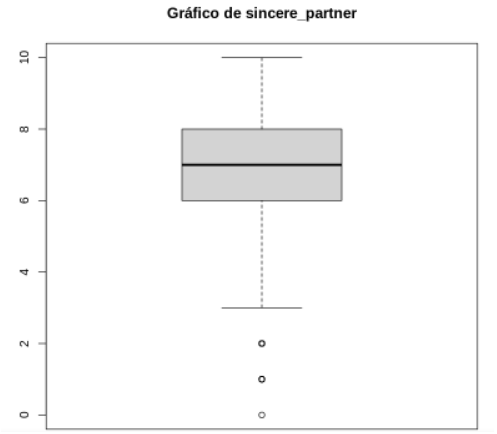
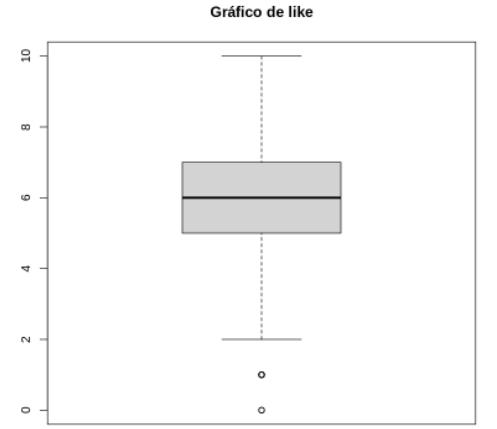
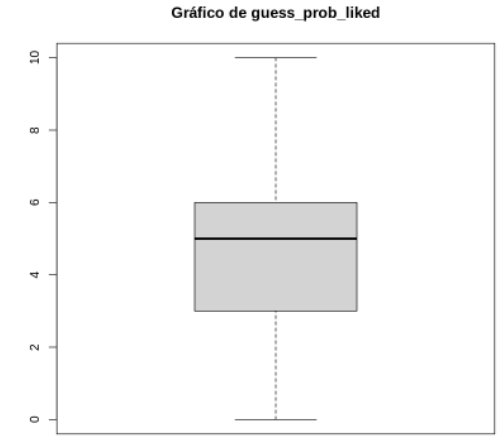


Gráfico de attractive_partner
10
8
6
4
2
0





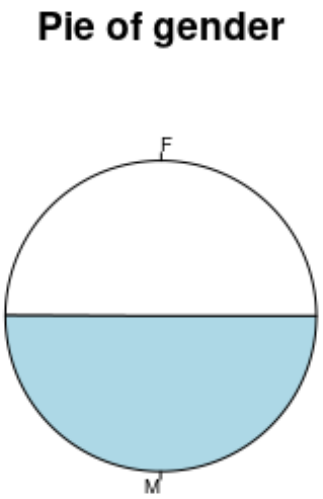


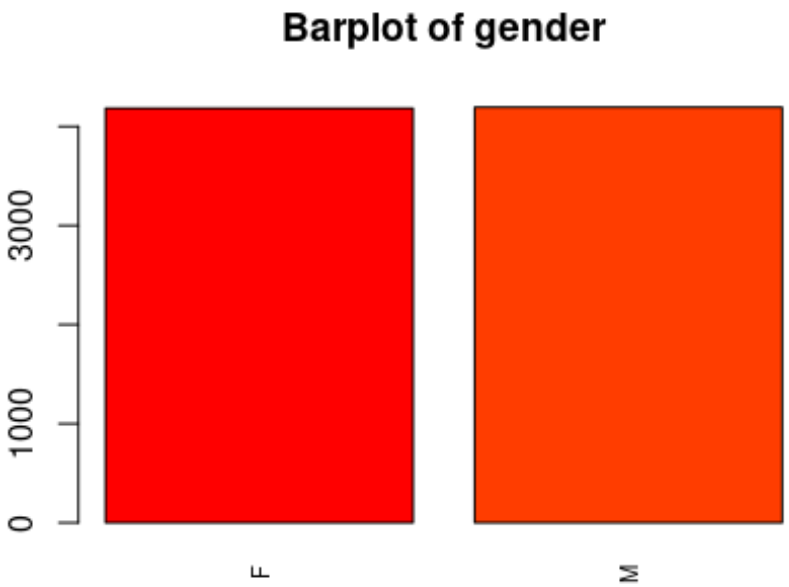


As we can see in the boxplots, some of the variables have outliers, but those are in the expected interval and can be possible. For this reason we won't delete those values because if we do it, we are losing information of possible values.

Descriptive analysis of modified variables

Variable 1: Gender



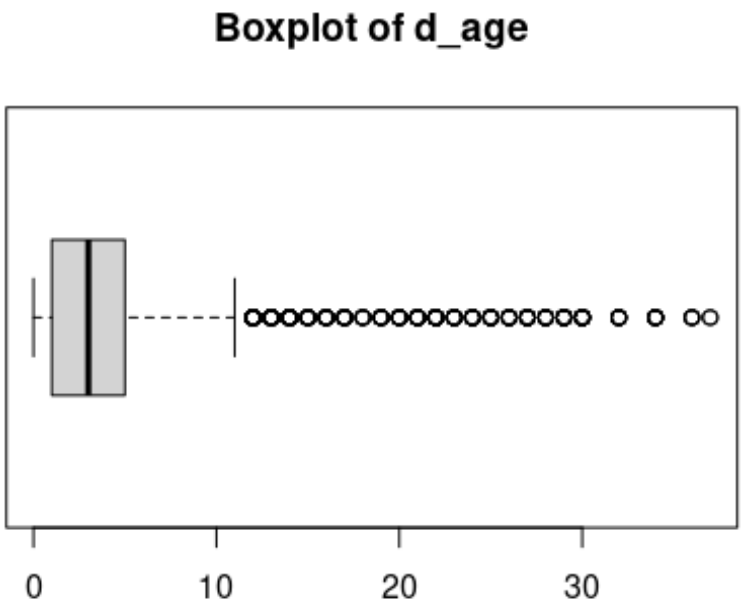


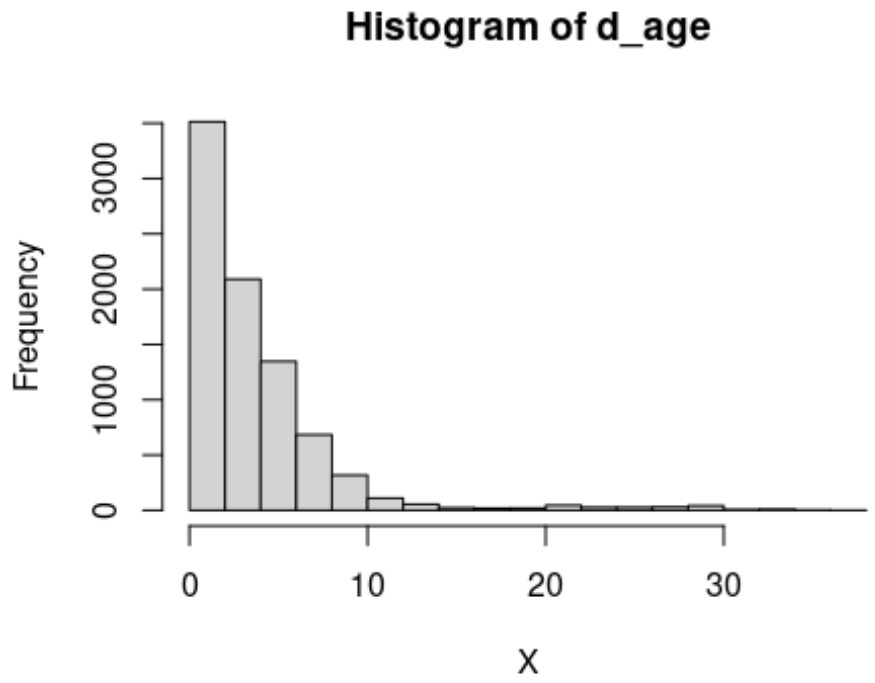
Modalities: 2

Modalities	Frequency	Proportions
F	*4184*	*0.4994032*
M	*4194*	*0.5005968*

From the frequency table, we can see that our sample is very balanced since we have almost exactly 50% men and women. This is great since we want our sample of subjects to be representative of the population so that our conclusions aren't skewed by having an unrepresentative sample.

Variable 2: D_age

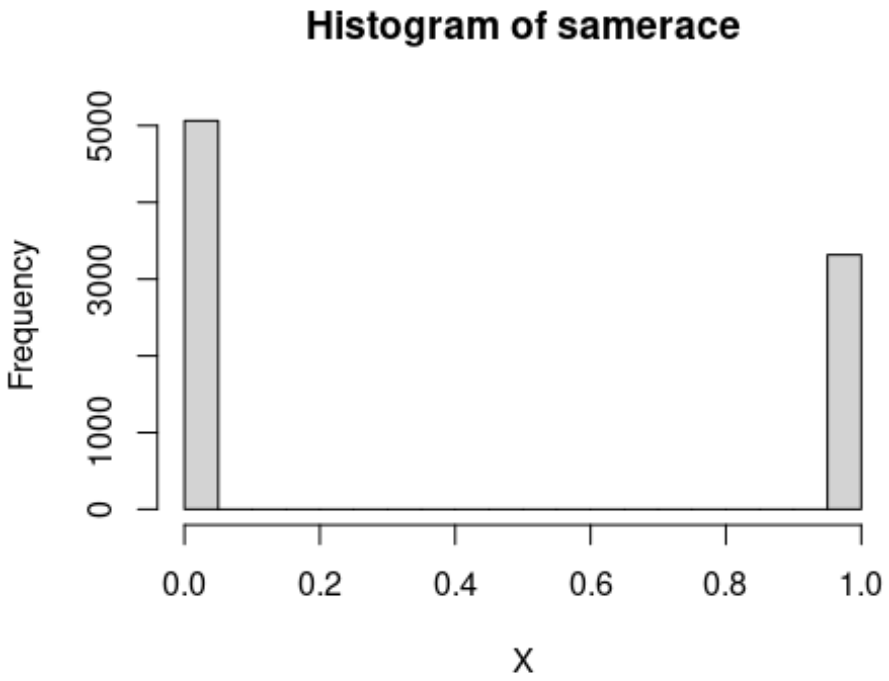


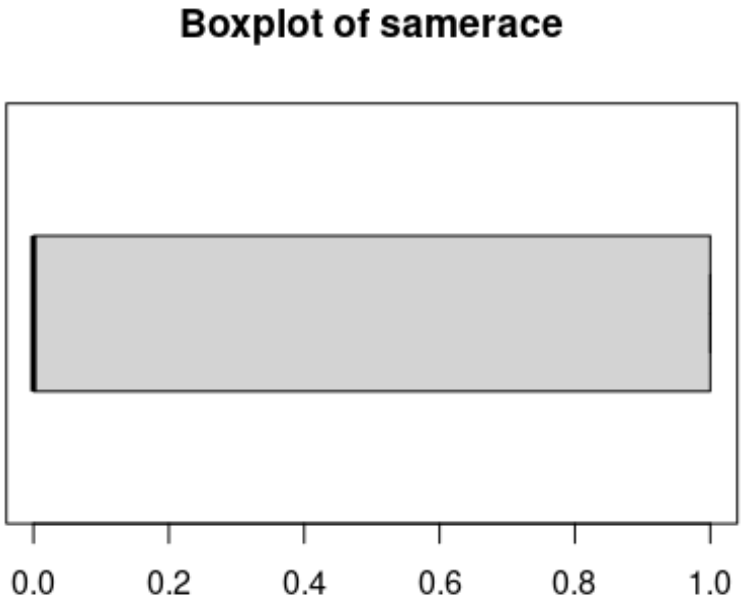


Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
D_age	0.000	1.000	3.000	4.186	5.000	37.000	4.5962	1.0981

From the frequency table, we can see that 50% of the respondents prefer having a partner within 3 years of their age in range. We can also see that 75% of the respondents prefer having a partner within 5 years of their age range. From the histogram, we can also clearly see that almost every respondent wanted an age difference of less than 10 years. We can deduce from that that people almost always prefer having partners close in age.

Variable 3: Samerace

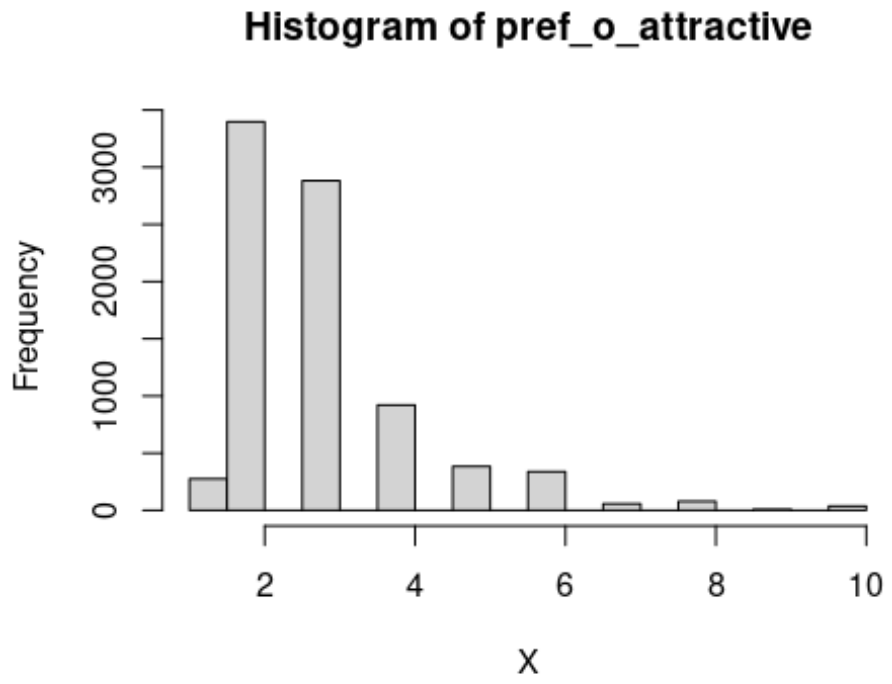


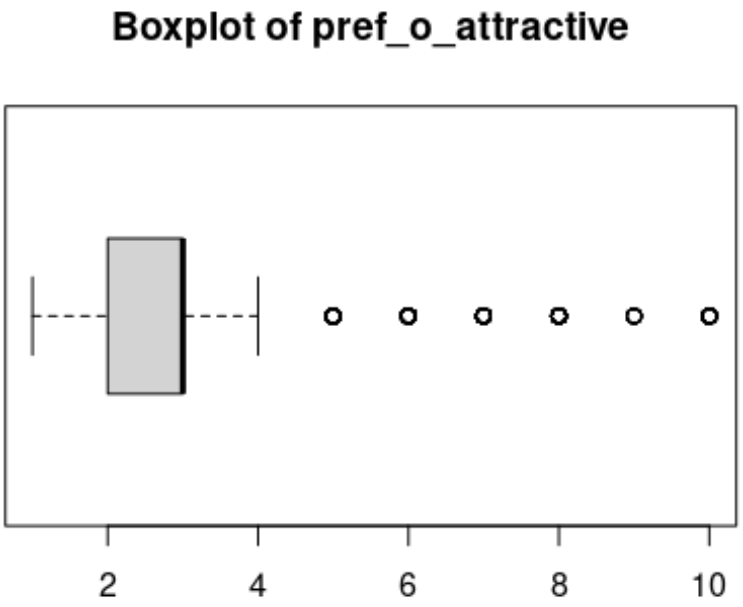


Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
Samerace	0.0000	0.0000	0.0000	0.3958	1.0000	1.0000	0.4891	1.2356

From the histogram, we can see that almost 60% of the respondents answered that they didn't have the same race as their partner, which we could say is good, since it means we have a good variability of races in our sample of the population and that our sample is diverse, when it comes to race. We should not have biases in our conclusions in relation to race because of that.

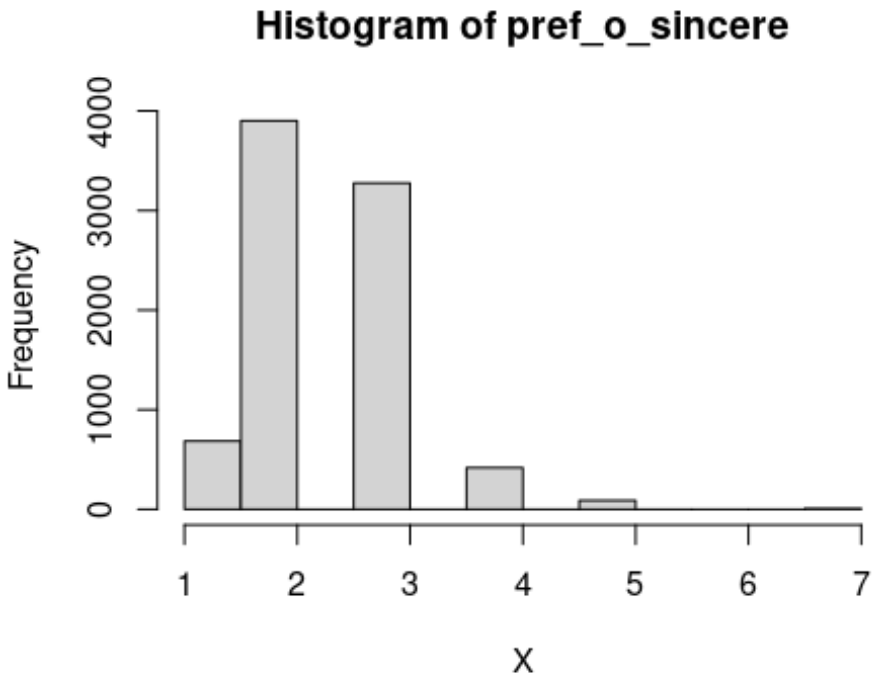
Variable 4: pref_o_attractive

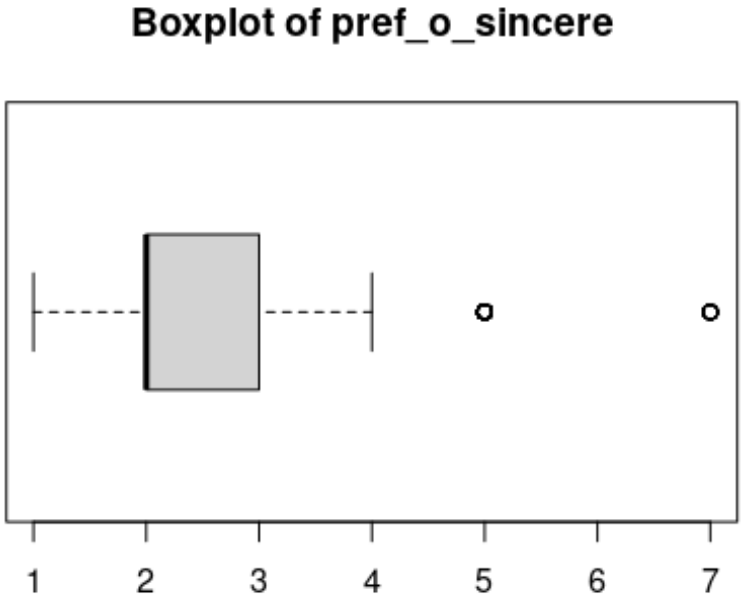




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
pref_o_attractive	1.000	2.000	3.000	2.963	3.000	10.000	1.3374	0.4514

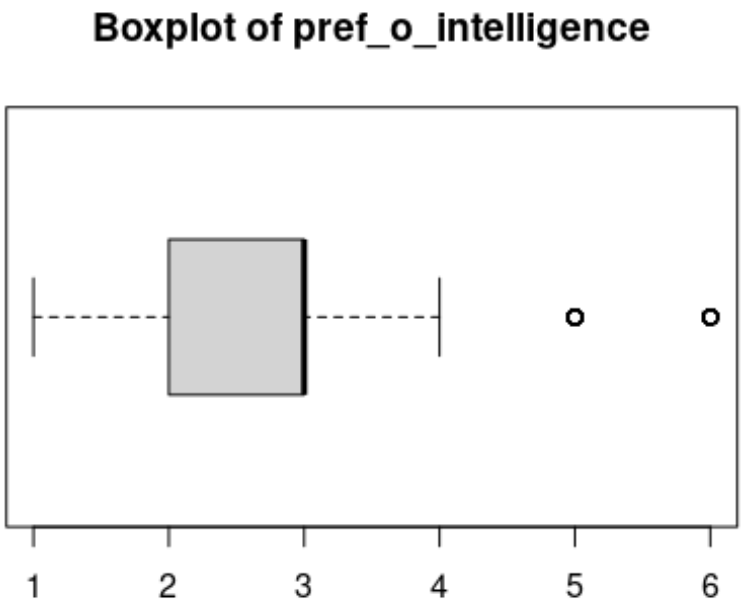
Variable 5: pref_o_sincere

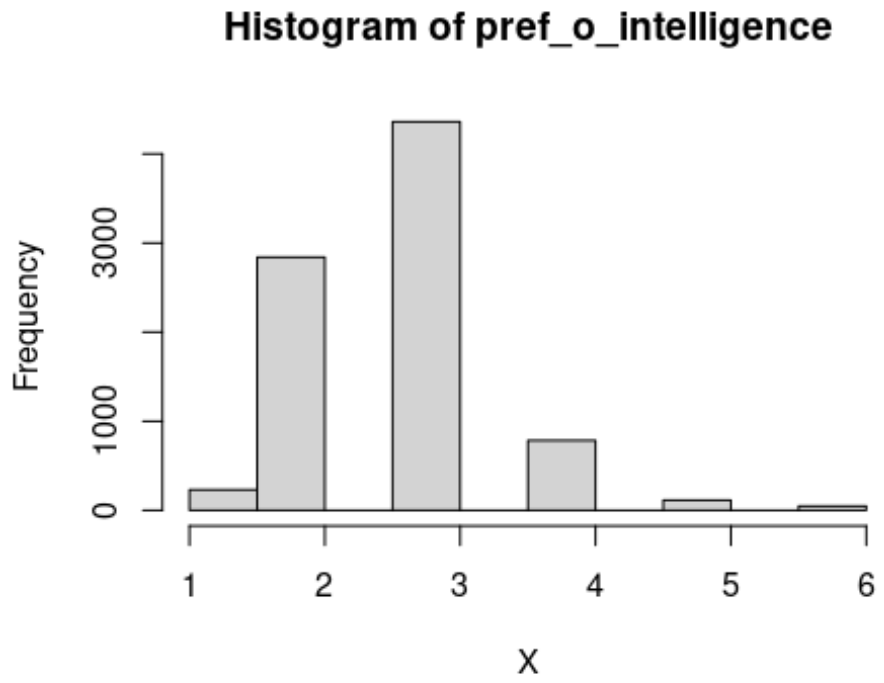




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
pref_o_sincere	1.000	2.000	2.000	2.446	3.000	7.000	0.7718	0.3155

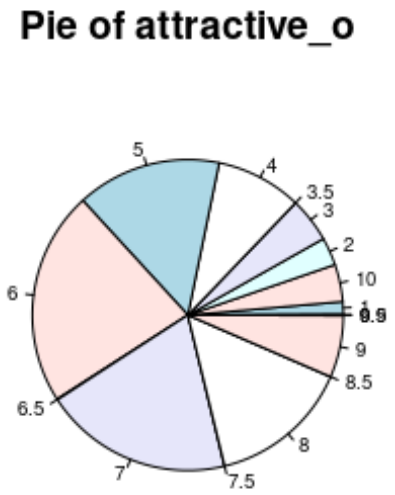
Variable 6: pref_o_intelligence

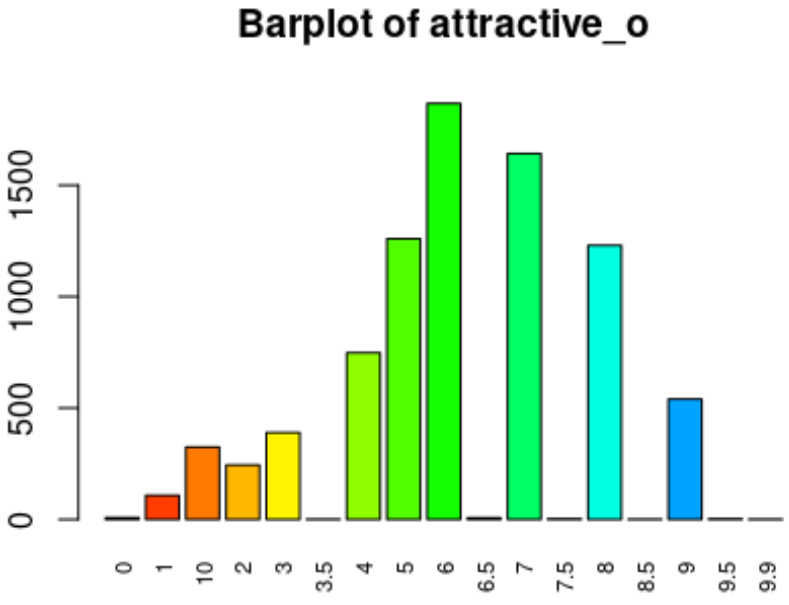




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
pref_o_intelligence	*1.000*	*2.000*	*3.000*	*2.744*	*3.000*	*6.000*	*0.7638*	*0.2784*

Variable 7: attractive_o



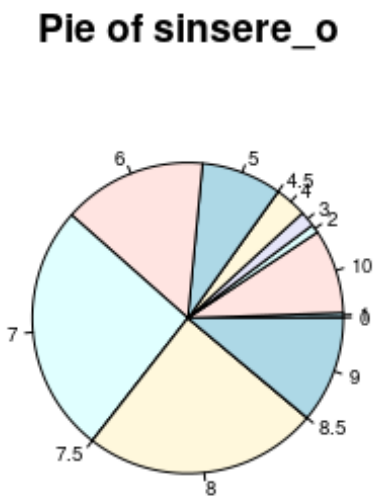


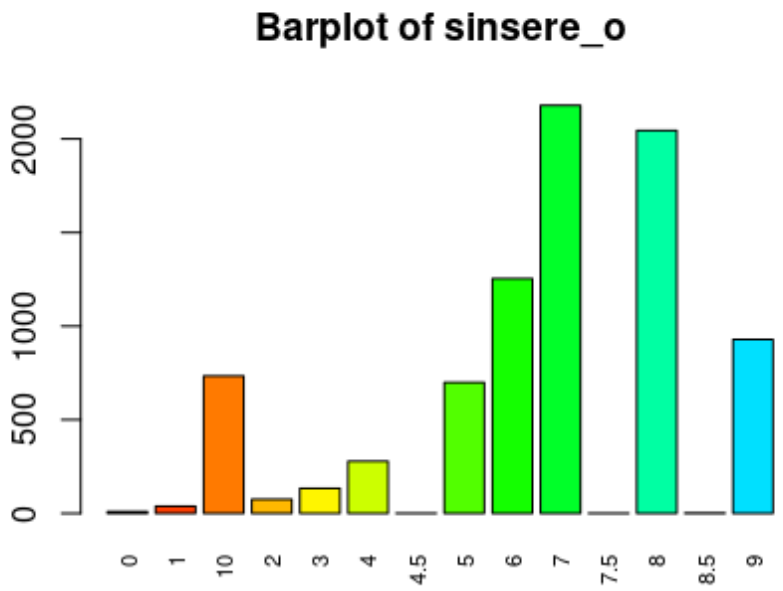
Modalities: 17

Modalities	Frequency	Proportions
0	*8*	*0.0009548818*
1	*108*	*0.0128909048*
2	*244*	*0.0291238959*
3	*390*	*0.0465504894*
3.5	*1*	*0.0001193602*
4	*748*	*0.0892814514*
5	*1260*	*0.1503938888*
6	*1867*	*0.2228455479*
6.5	*7*	*0.0008355216*
7	*1642*	*0.1959894963*
7.5	*3*	*0.0003580807*
8	*1230*	*0.1468130819*
8.5	*1*	*0.0001193602*
9	*540*	*0.0644545238*

9.5	*1*	*0.0003580807*
9.9	*1*	*0.0001193602*
10	*325*	*0.0387920745*

Variable 8: sincere_o



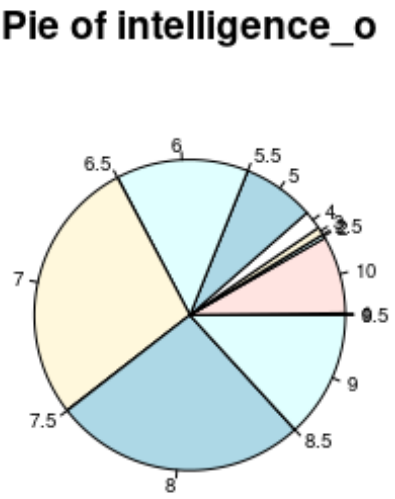


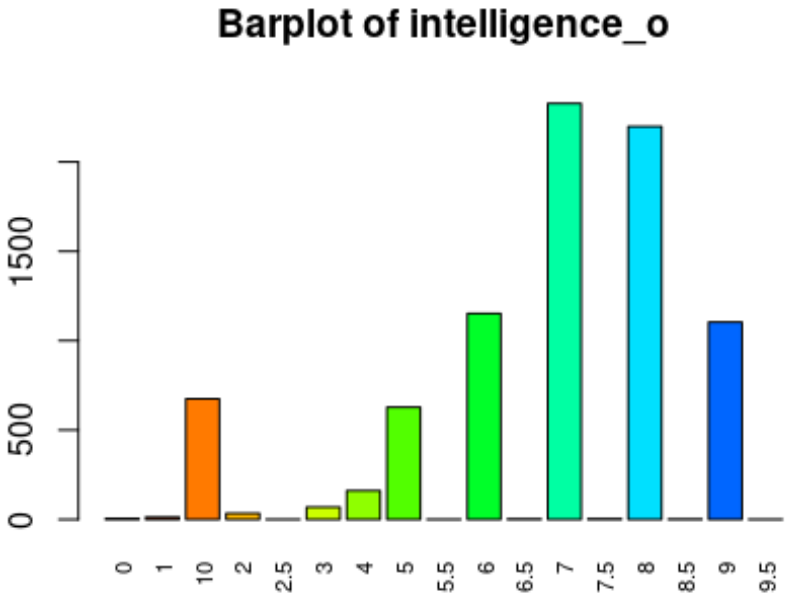
Modalities: 14

Modalities	Frequency	Proportions
0	*9*	*0.0010742421*
1	*38*	*0.0045356887*
2	*75*	*0.0089520172*
3	*134*	*0.0159942707*
4	*278*	*0.0331821437*
4.5	*1*	*0.0001193602*

5	*699*	*0.0834328002*
6	*1254*	*0.1496777274*
7	*2179*	*0.2600859394*
7.5	*1*	*0.0001193602*
8	*2045*	*0.2440916687*
8.5	*2*	*0.0002387205*
9	*929*	*0.1108856529*
10	*734*	*0.0876104082*

Variable 9: intelligence_o



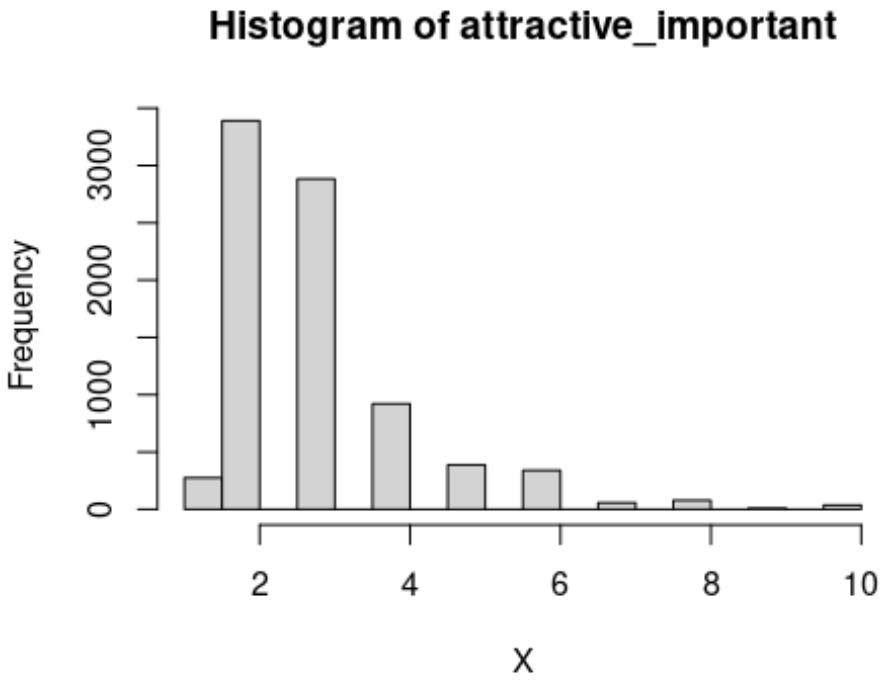


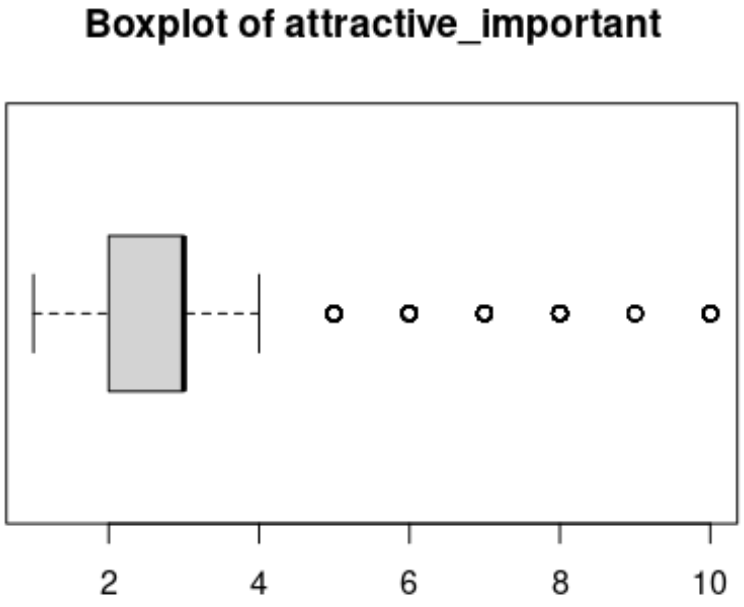
Number of modalities: 17

Modalities	Frequency	Proportions
0	*5*	*0.0005968011*
1	*13*	*0.0015516830*
2	*34*	*0.0040582478*
2.5	*1*	*0.0001193602*
3	*69*	*0.0082358558*
4	*161*	*0.0192169969*
5	*628*	*0.0749582239*
5.5	*1*	*0.0001193602*
6	*1152*	*0.1375029840*
6.5	*3*	*0.0003580807*
7	*2327*	*0.2777512533*
7.5	*4*	*0.0004774409*
8	*2198*	*0.2623537837*
8.5	*2*	*0.0002387205*

9	*1104*	*0.1317736930*
9.5	*1*	*0.0001193602*
10	*675*	*0.0805681547*

Variable 10: attractive_important

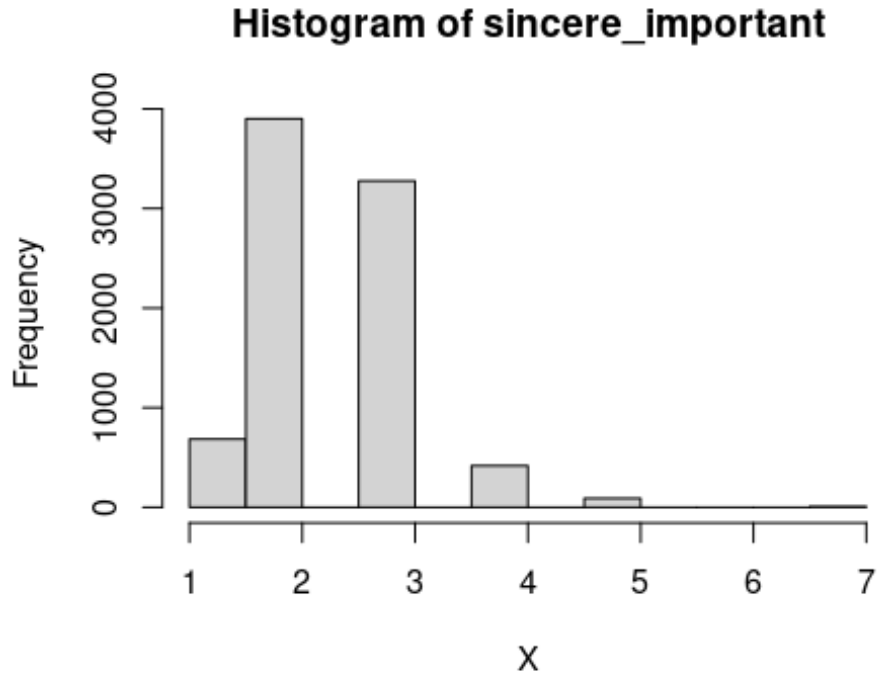


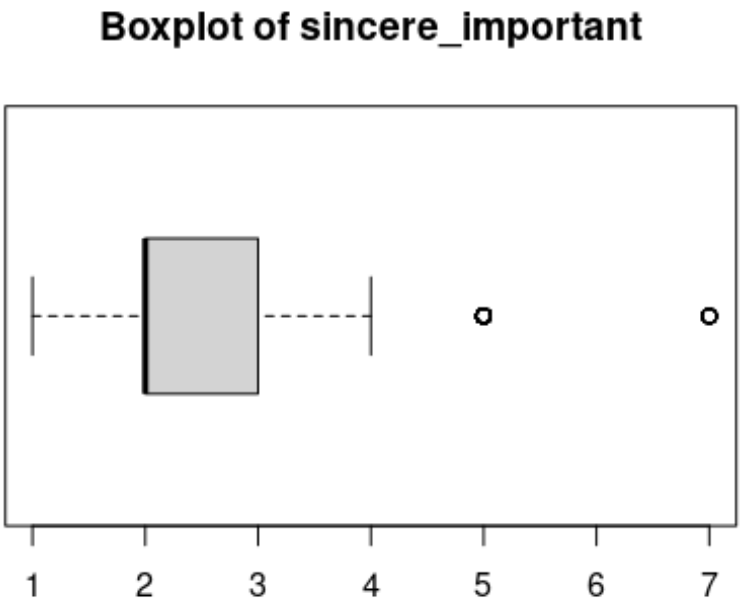


Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
attractive_important	1.0000	2.0000	3.0000	2.966	3.000	10.000	1.3399	0.4518

From the boxplot we can see that 75% of respondents attribute a 3/10 for attractiveness and that there are only outliers that attribute it more than 4. We can then conclude that almost all our respondents don't think attractiveness is important in their partner.

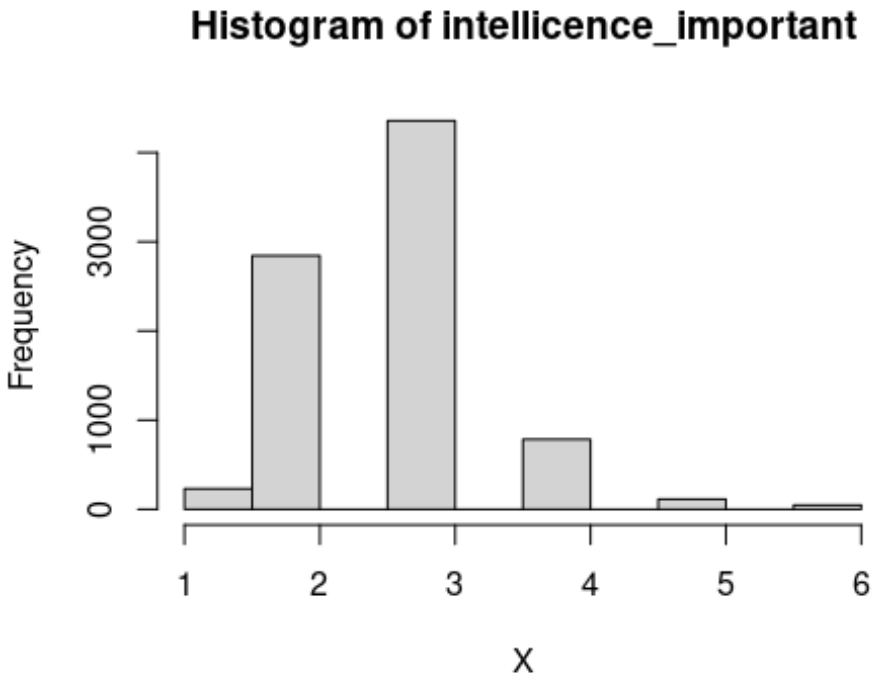
Variable 11: sincere_important

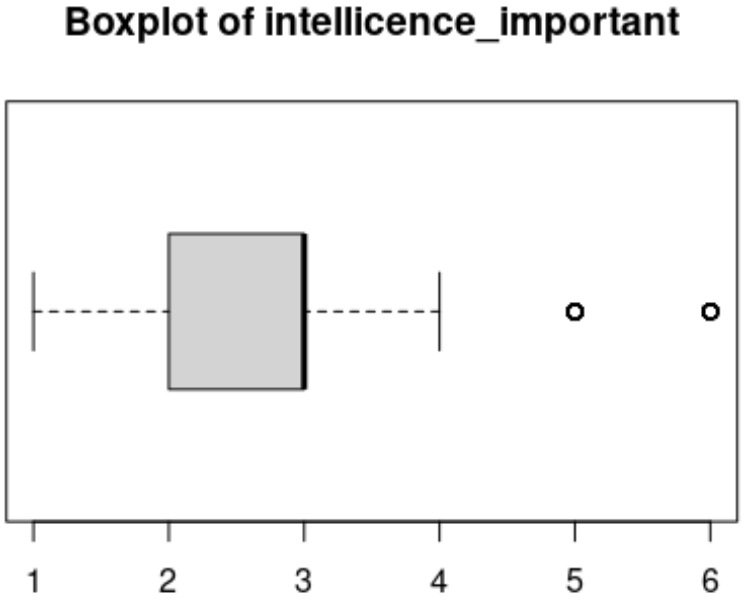




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
sincere_important	1.0000	2.0000	2.0000	2.447	3.000	7.000	0.7723	0.3157

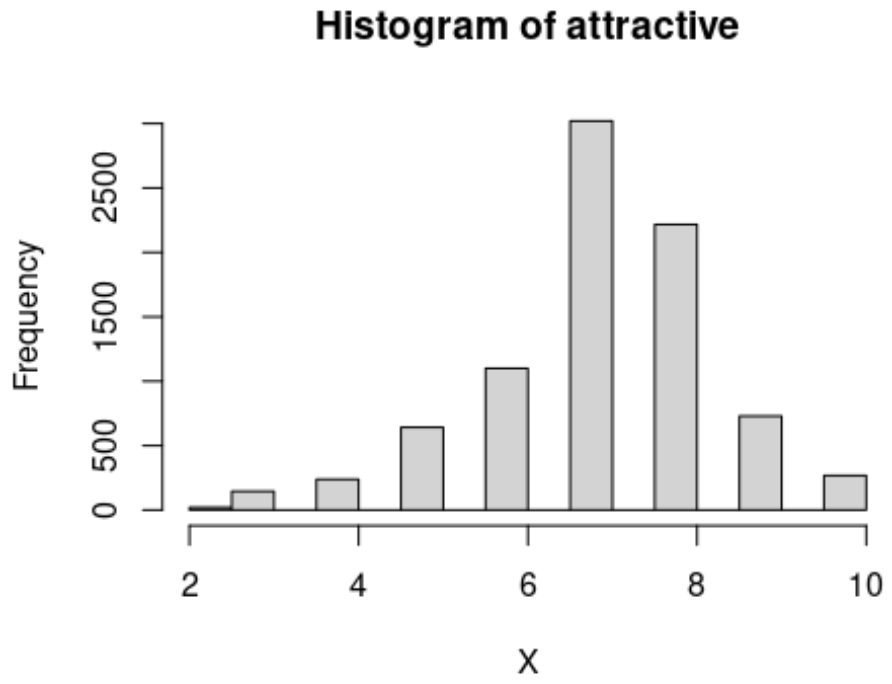
Variable 12: intelligence_important

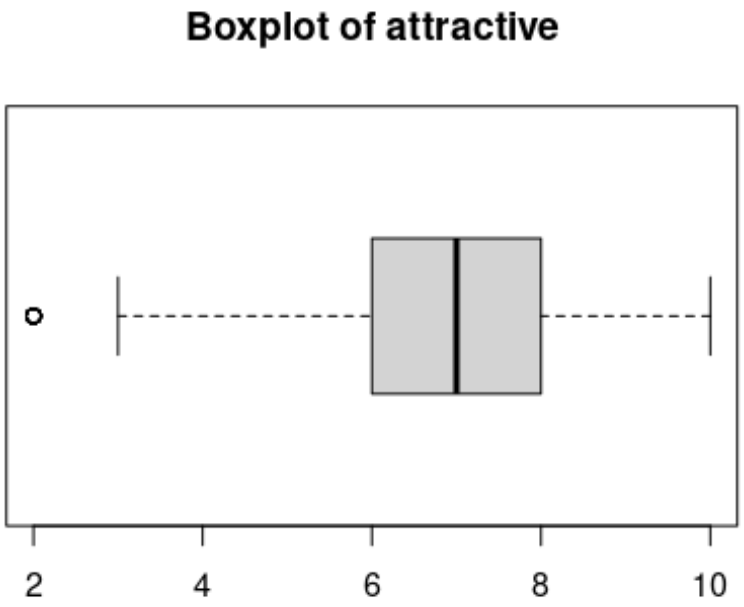




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
intelligence_important	1.0000	2.0000	3.0000	2.743	3.000	6.000	0.7639	0.2785

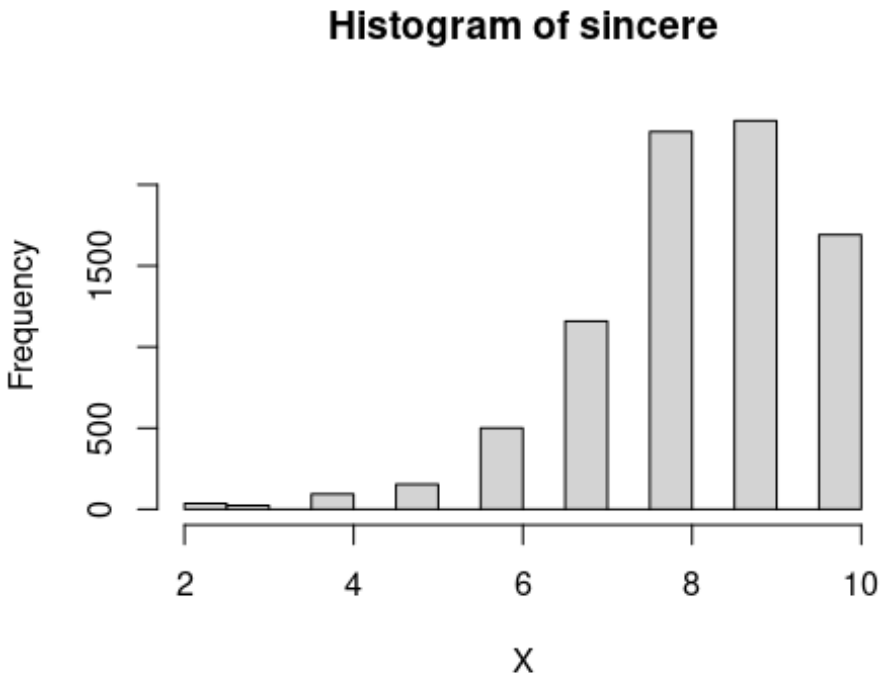
Variable 13: attractive

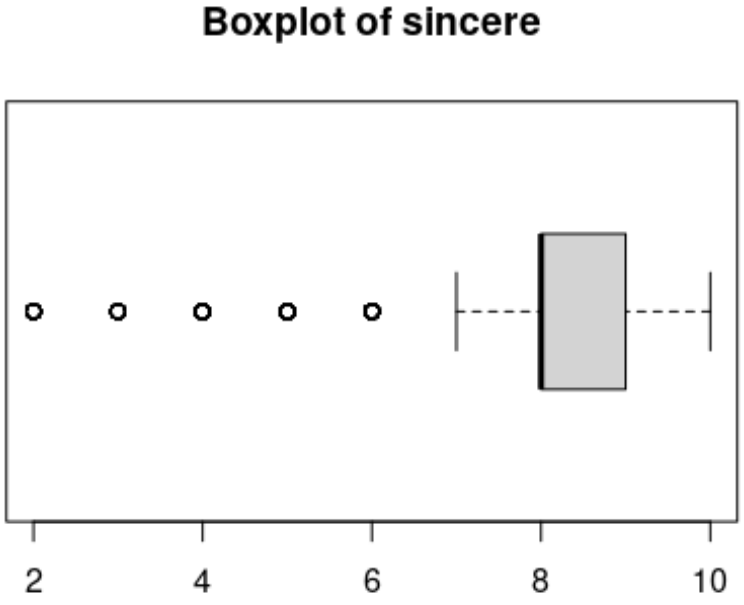




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
attractive	2.0000	6.0000	7.0000	7.084	8.000	10.000	1.3870	0.1958

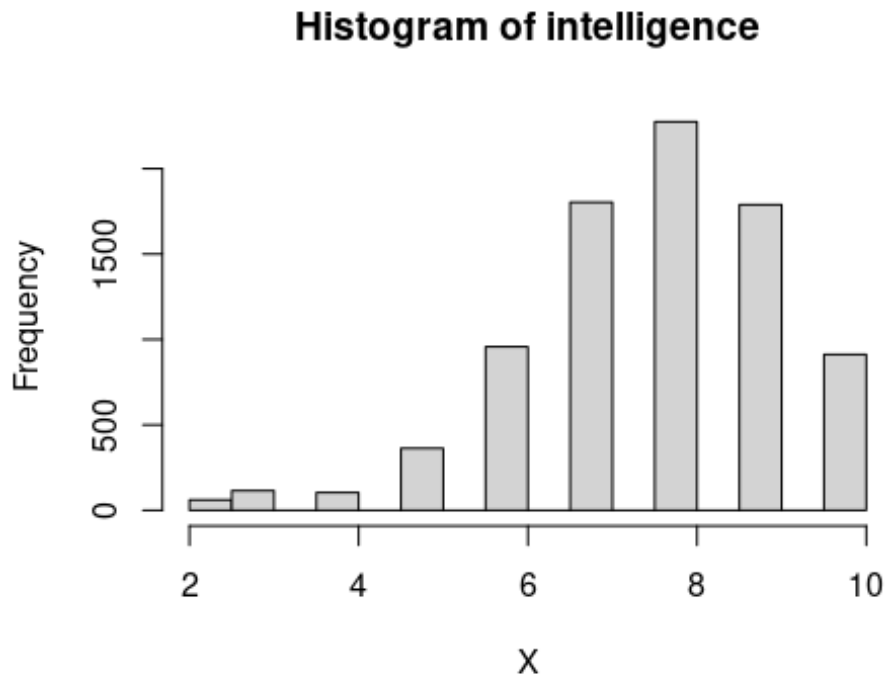
Variable 14: sincere

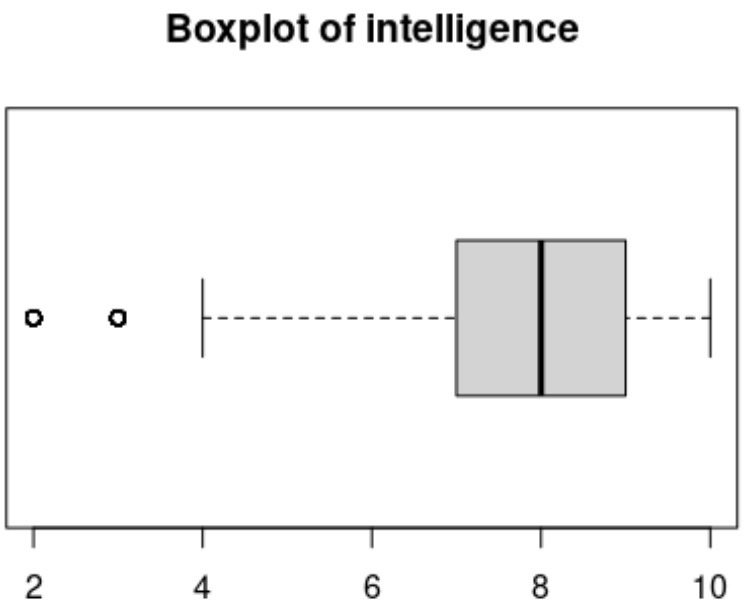




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
sincere	2.0000	8.000	8.000	8.291	9.000	10.000	1.3990	0.1687

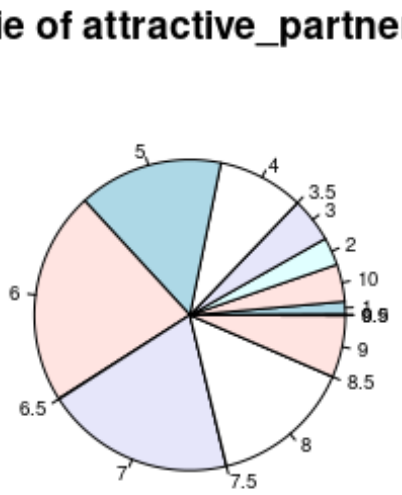
Variable 15: intelligence

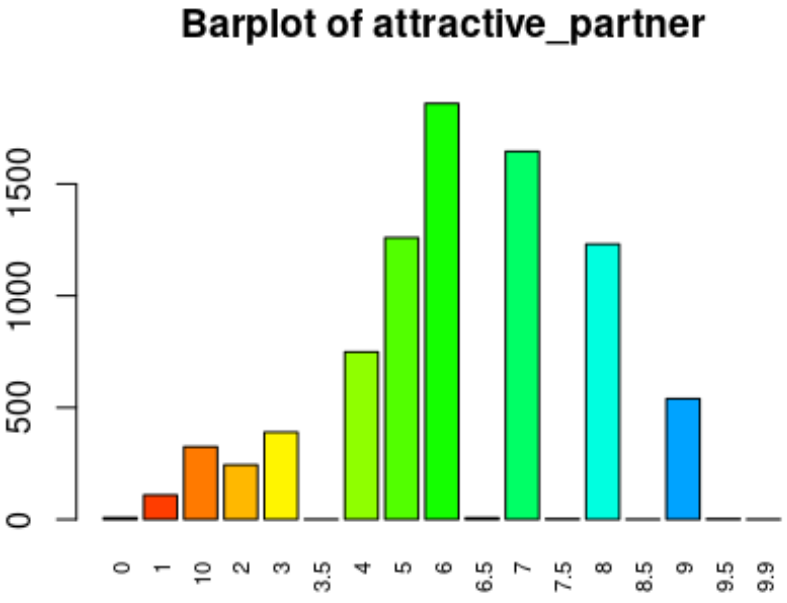




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
intelligence	*2.0000*	*7.000*	*8.000*	*7.696*	*9.000*	*10.000*	*1.5565*	*0.2023*

Variable 16: attractive_partner





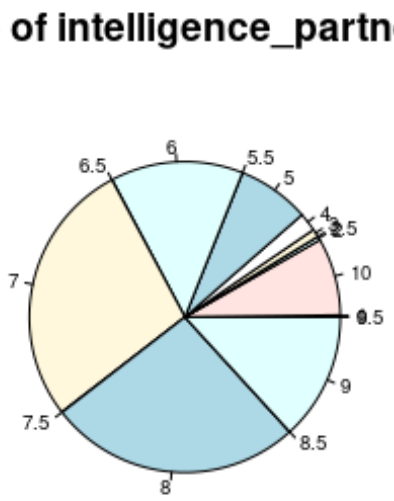
Modalities: 17

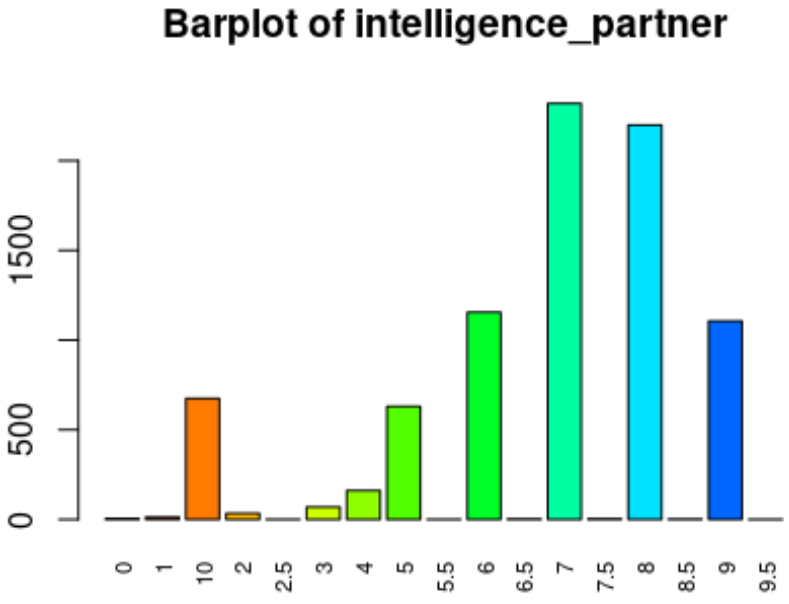
Modalities	Frequency	Proportions
0	*8*	*0.0009548818*
1	*109*	*0.0130102650*
2	*244*	*0.0291238959*
3	*390*	*0.0465504894*
3.5	*1*	*0.0001193602*
4	*749*	*0.0894008116*
5	*1260*	*0.1503938888*
6	*1860*	*0.2220100263*
6.5	*7*	*0.0008355216*
7	*1646*	*0.1964669372*
7.5	*3*	*0.0003580807*
8	*1231*	*0.1469324421*
8.5	*1*	*0.0001193602*
9	*540*	*0.0644545238*

9.5	*3*	*0.0003580807*
9.9	*1*	*0.0001193602*
10	*325*	*0.0387920745*

From the Venn diagram, we can easily see that around 60% of the respondants were rated 6 and above which means we have a normal distribution here as well, but if we assume 5 as the average person, our distribution is skewed towards attractiveness by a margin. Even though we would need to have more data to know the explanation of this, we could postulate the hypothesis that we either have prettier people than average in our sample, that people might have assumed 6 was the average, which would explain why it is the biggest group or it is for other reasons we don’t know.

Variable 17: intelligence_partner





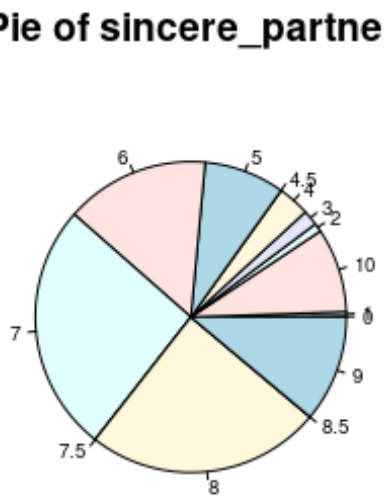
Modalities: 17

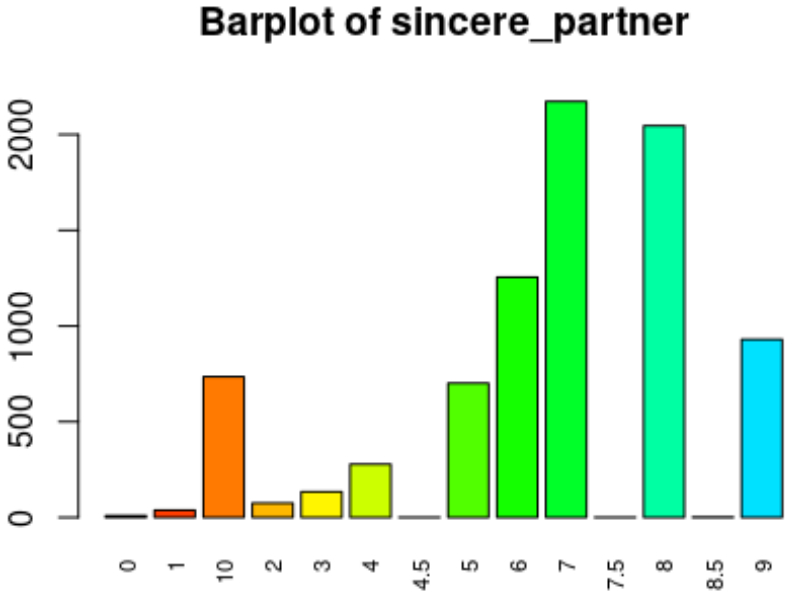
Modalities	Frequency	Proportions
0	*5*	*0.0005968011*
1	*13*	*0.0015516830*
2	*34*	*0.0040582478*
2.5	*1*	*0.0001193602*
3	*69*	*0.0082358558*
4	*161*	*0.0192169969*
5	*630*	*0.0751969444*
5.5	*1*	*0.0001193602*
6	*1155*	*0.1378610647*
6.5	*3*	*0.0003580807*
7	*2319*	*0.2767963714*
7.5	*4*	*0.0004774409*
8	*2199*	*0.2624731439*
8.5	*2*	*0.0002387205*

9	*1106*	*0.1320124135*
9.5	*1*	*0.0001193602*
10	*675*	*0.0805681547*

From the Venn diagram, we can easily see that people have rated their partner's intelligence very high since almost 90% of the answers are between 5 and 10 and we know that intelligence in general follows a normal distribution centered around 100IQ and it should be symmetric, which it is not so much. We either have a group of respondents that are very intelligent or respondents were nice when attributing scores. The explanation might also be something else, we cannot come to conclusions about that with the data we have.

Variable 18: sincere_partner

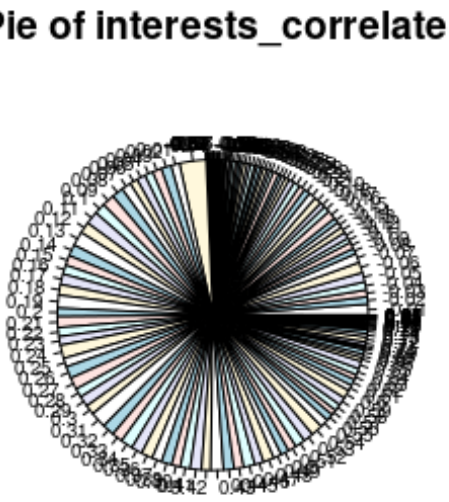


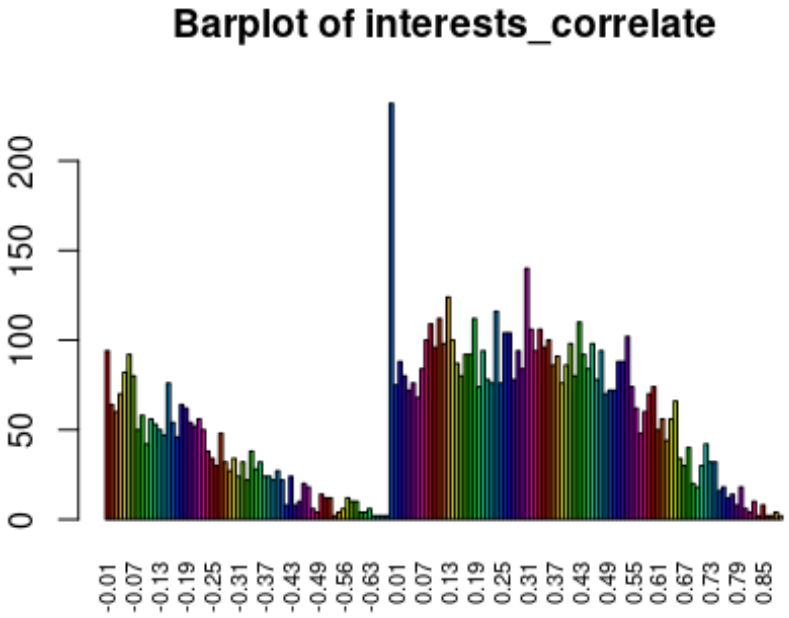


Modalities: 14

Modalities	Frequency	Proportions
0	*9*	*0.0010742421*
1	*38*	*0.0045356887*
2	*75*	*0.0089520172*
3	*134*	*0.0159942707*
4	*278*	*0.0331821437*
4.5	*1*	*0.0001193602*
5	*701*	*0.0836715206*
6	*1255*	*0.1497970876*
7	*2173*	*0.2593697780*
7.5	*1*	*0.0001193602*
8	*2046*	*0.2442110289*
8.5	*2*	*0.0002387205*
9	*930*	*0.1110050131*
10	*735*	*0.0877297684*

Variable 19: interests_correlate

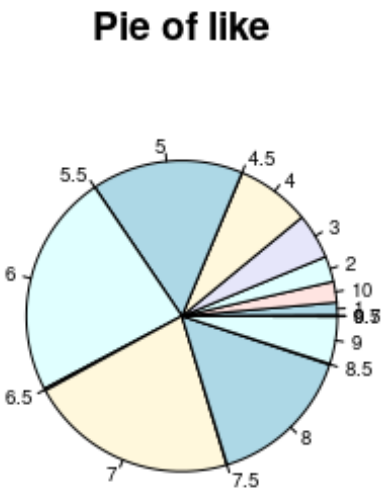


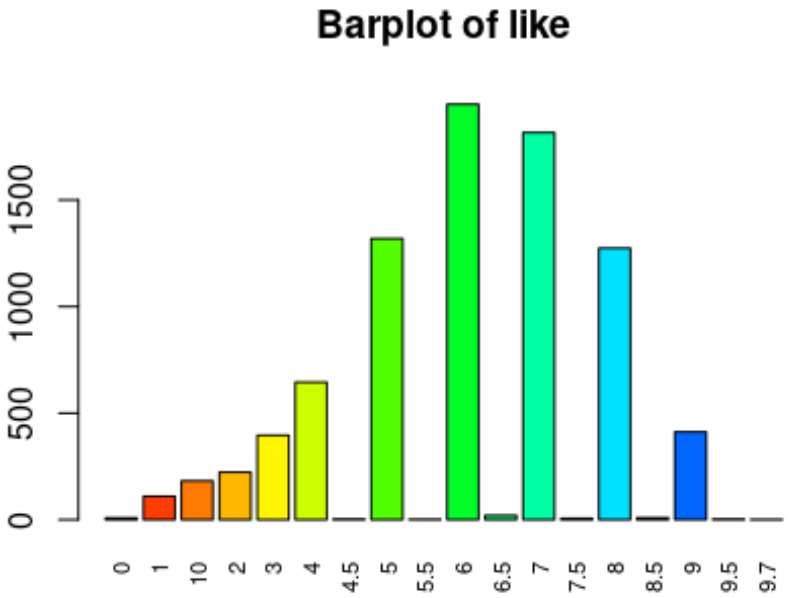


Here we have 155 modalities and hence why the table of frequencies and proportions is simply too big to be of importance. However, we can see from the barplot that if the graduation was done correctly in the x axis, we would have a normal distribution. This is

good to know since if we didn't have that, we could think that our sample might not have been representative of the population.

Variable 20: like



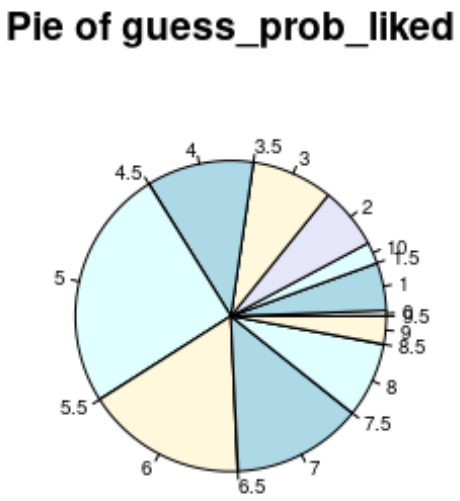


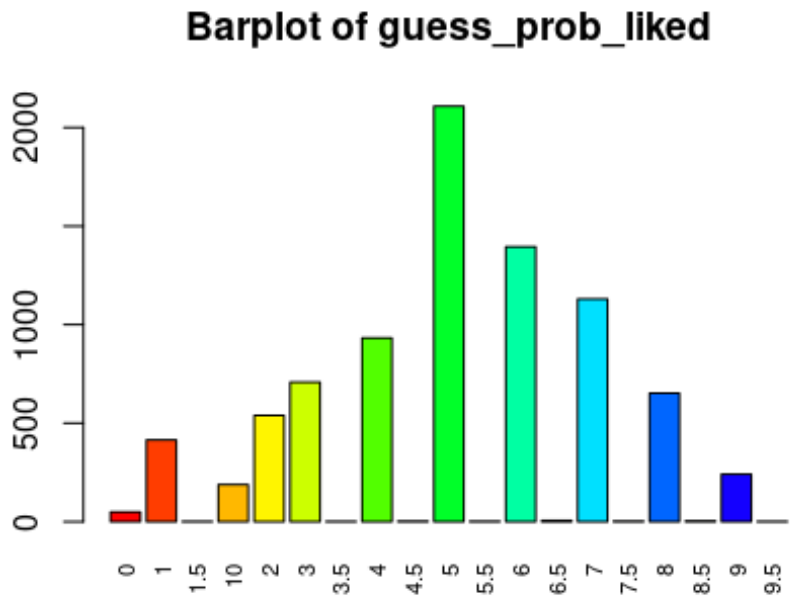
Modalities: 18

Modalities	Frequency	Proportions
0	*8*	*0.0009548818*
1	*110*	*0.0131296252*
2	*223*	*0.0266173311*
3	*396*	*0.0472666508*
4	*645*	*0.0769873478*
4.5	*3*	*0.0003580807*
5	*1319*	*0.1574361423*
5.5	*2*	*0.0002387205*
6	*1949*	*0.2326330867*
6.5	*20*	*0.0023872046*
7	*1816*	*0.2167581762*
7.5	*6*	*0.0007161614*

8	*1274*	*0.1520649320*
8.5	*9*	*0.0010742421*
9	*412*	*0.0491764144*
9.5	*3*	*0.0003580807*
9.7	*1*	*0.0001193602*
10	*182*	*0.0217235617*

Variable 20: like



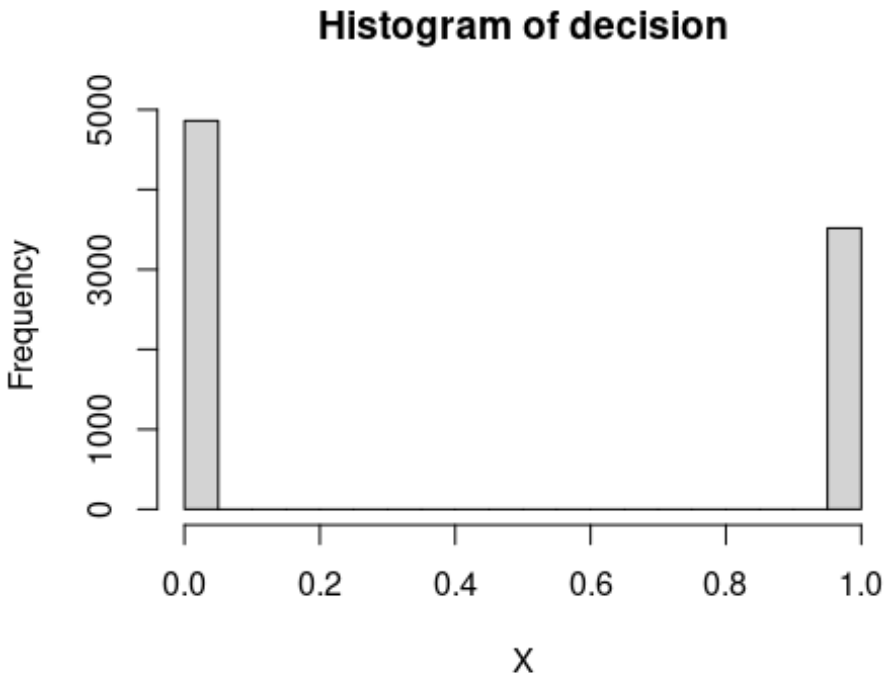


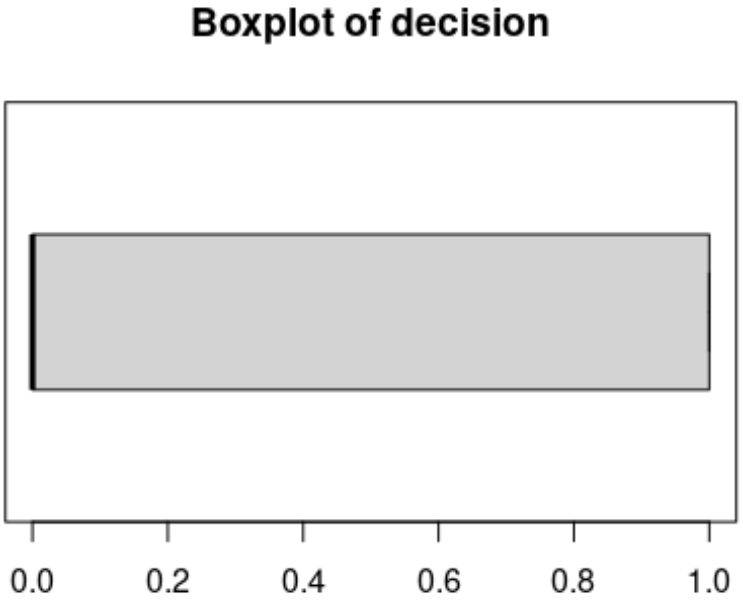
Modalities: 19

Modalities	Frequency	Proportions
0	*49*	*0.0058486512*
1	*415*	*0.0495344951*
1.5	*1*	*0.0001193602*
2	*539*	*0.0643351635*
3	*708*	*0.0845070423*

3.5	1	*0.0001193602*
4	*932*	*0.1112437336*
4.5	*3*	*0.0003580807*
5	*2108*	*0.2516113631*
5.5	*2*	*0.0002387205*
6	*1395*	*0.1665075197*
6.5	*6*	*0.0007161614*
7	*1130*	*0.1348770590*
7.5	*3*	*0.0003580807*
8	*652*	*0.0778228694*
8.5	*4*	*0.0004774409*
9	*241*	*0.0287658152*
9.5	*1*	*0.0001193602*
10	*188*	*0.0224397231*

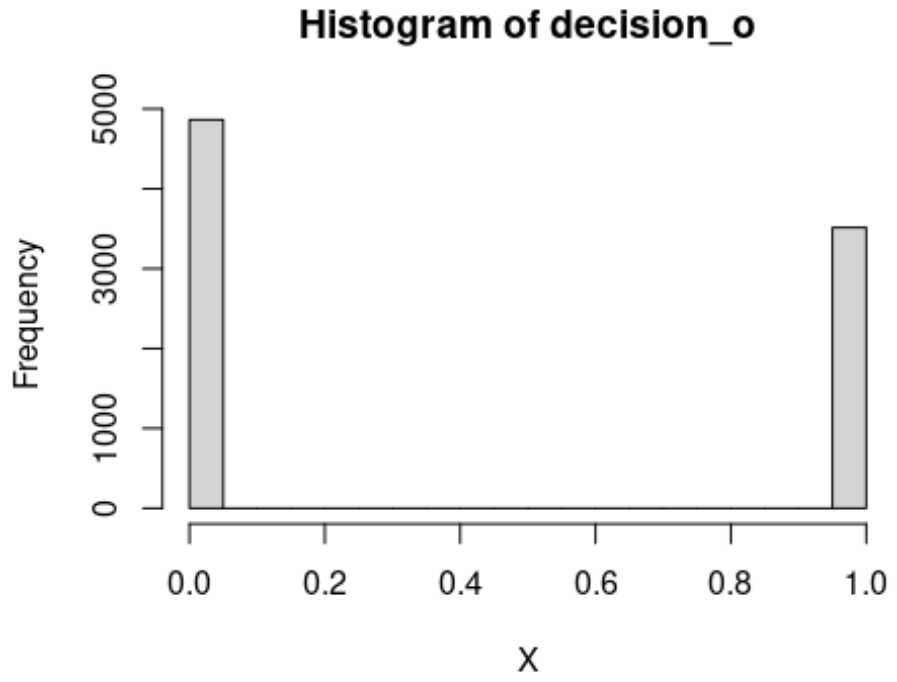
Variable 22: decision

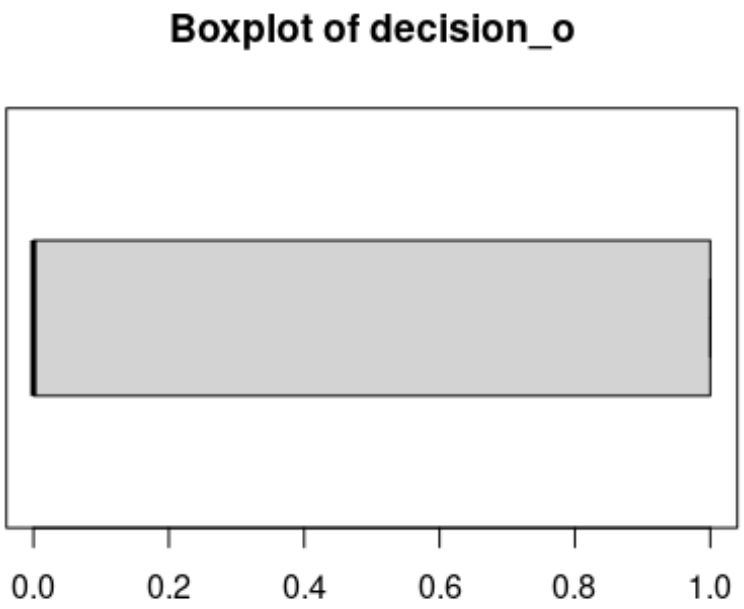




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
decision	*0.0000*	*0.000*	*0.000*	*0.4199*	*1.0000*	*1.0000*	*0.4936*	*1.1754*

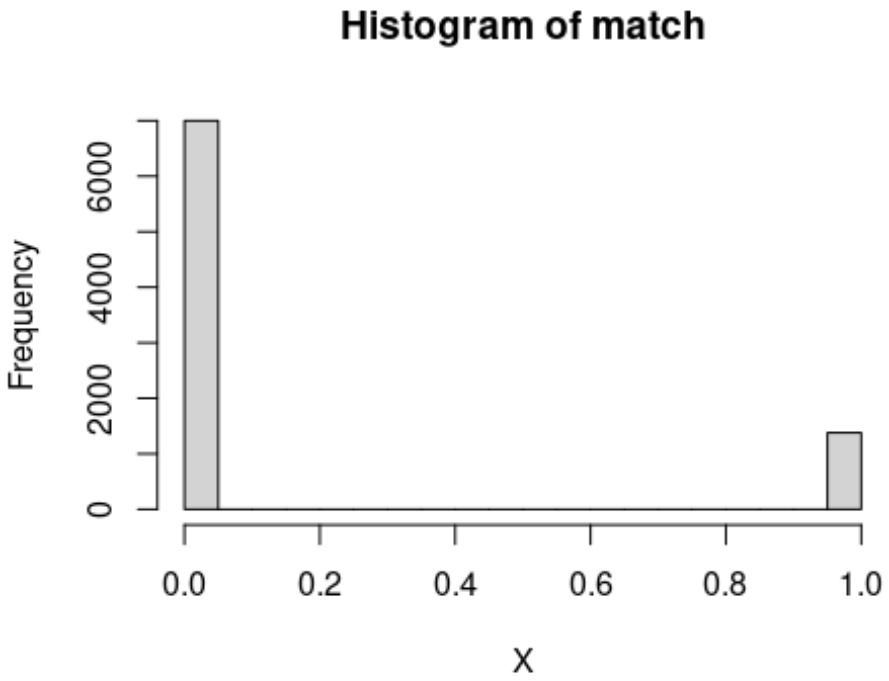
Variable 23: decision_o

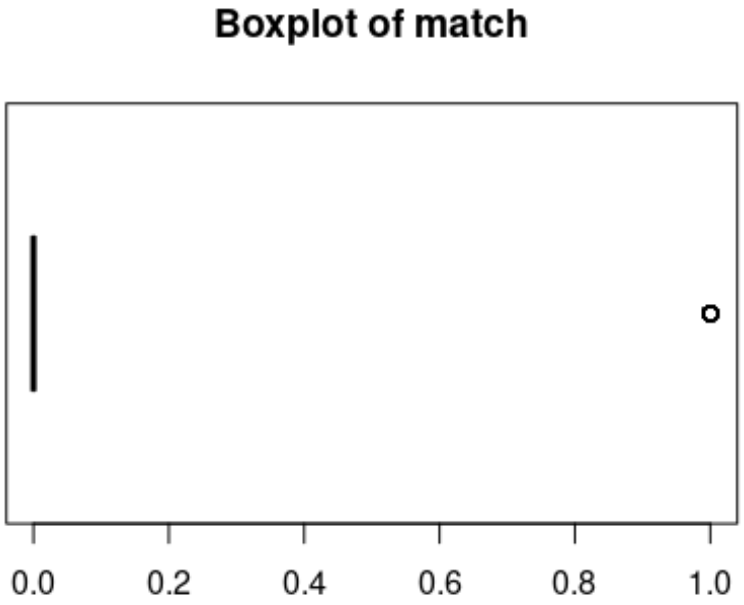




Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
Decision_o	0.0000	0.000	0.000	0.4196	1.0000	1.0000	0.4935	1.1763

Variable 23: match





Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	sd	vc
match	0.0000	0.000	0.000	0.1647	0.0000	1.0000	0.3709	2.2520

Descriptive analysis conclusions

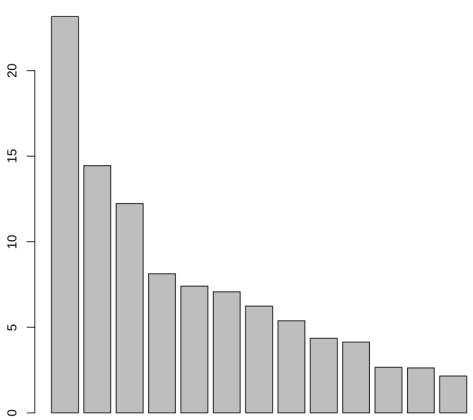
To conclude, we can see that our data and the sample of the population it comes from is very representative of the population which means that conclusions that can be drawn from it will not have any biases because of the sample.

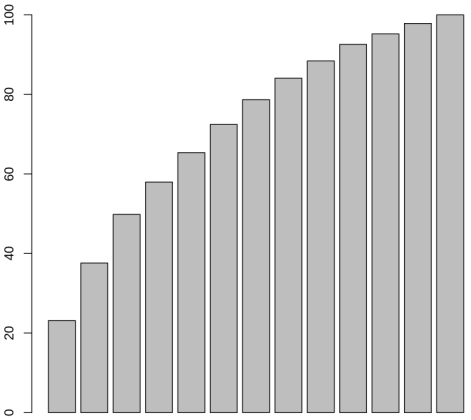
However, the methodology used to collect the data from the respondents limits us as to the conclusions we can draw from it since we cannot be sure if the answers of the respondents reflect their preferences in reality. Using a dataset from an application like Tinder or similar ones could be better if it provided us with the outcomes of the dates and relationships that came out of it. With that, we could look at the preferences people have and who they actually end up with and see if there are distortions in the preferences people think they have and their actual preferences when it comes to really being with someone.

PCA

In this section it is explained which steps have been followed for the development of the PCA and the results obtained.

First, we will plot the inertia of the different dimensions of our pca together with the plot of the accumulated inertia.

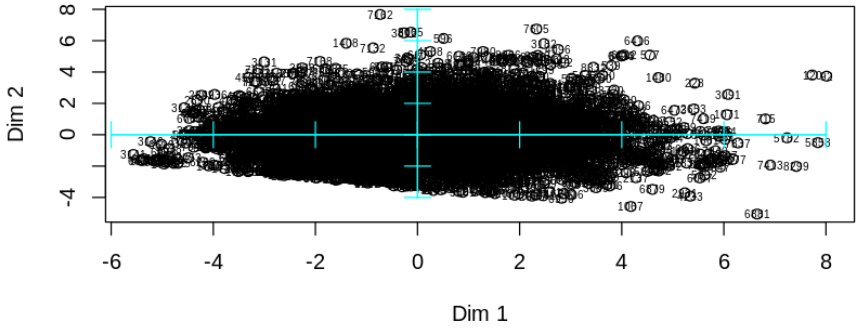




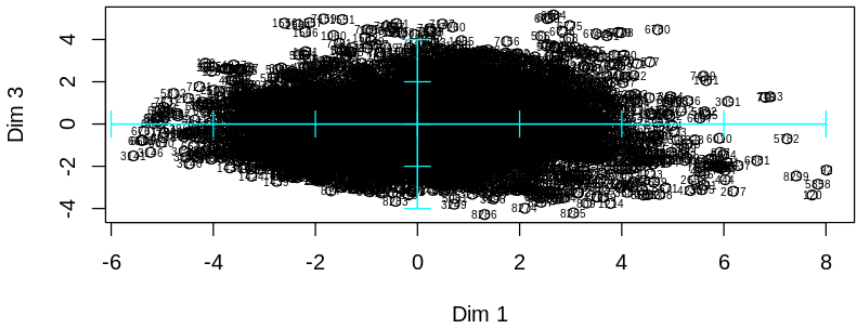
As we can see, it is in our interest to stay with 8 dimensions, as this way we remain with 80% of our inertia.

Now, we will plot every possible subspace to see which one are we working with.

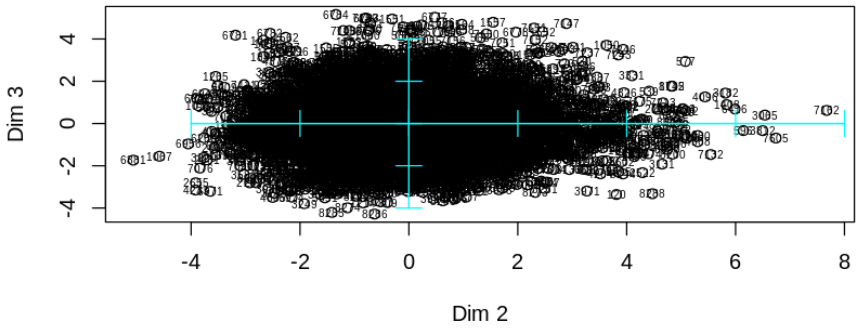
Subspace (X: 1 , Y: 2)



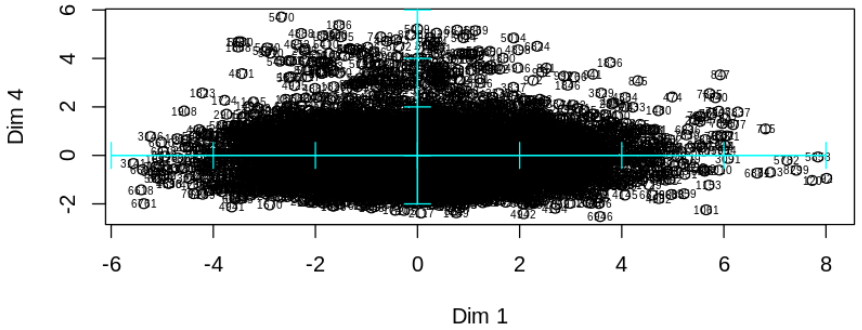
Subspace (X: 1 , Y: 3)



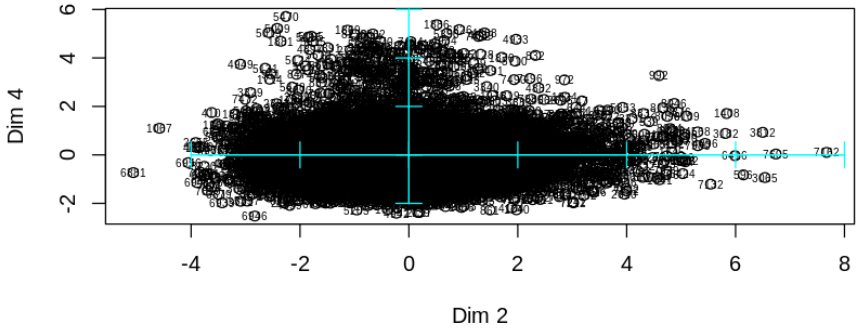
Subspace (X: 2 , Y: 3)



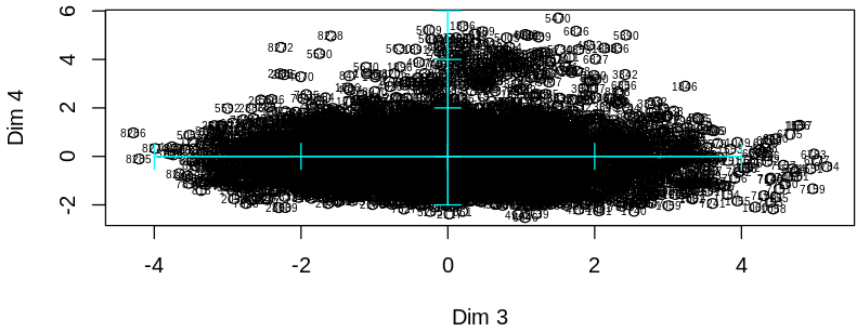
Subspace (X: 1 , Y: 4)



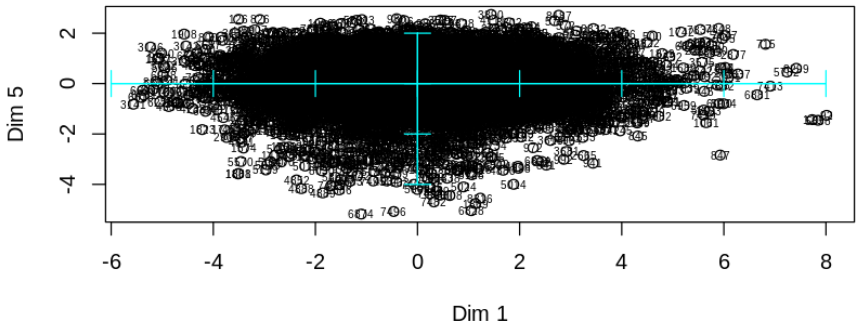
Subspace (X: 2 , Y: 4)



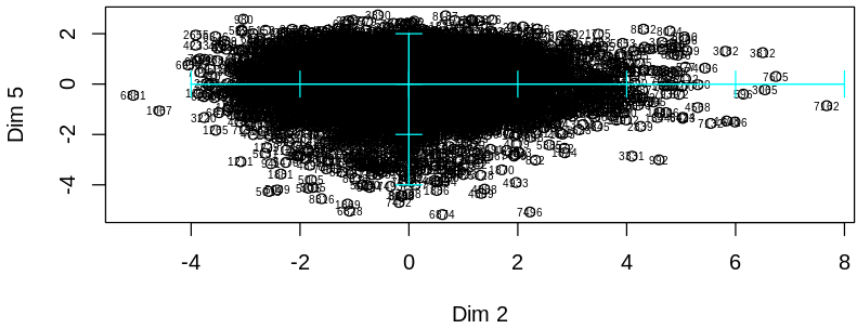
Subspace (X: 3 , Y: 4)



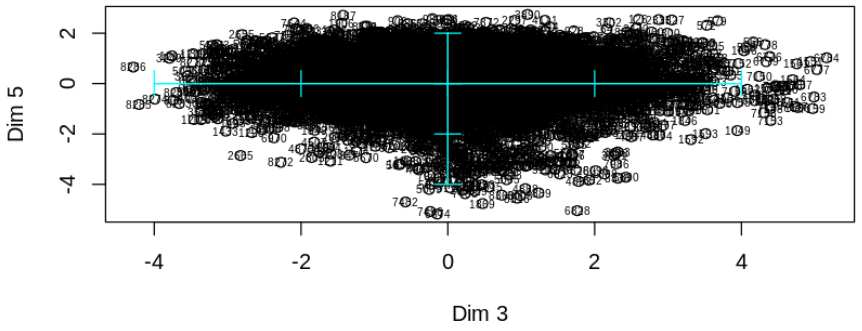
Subspace (X: 1 , Y: 5)



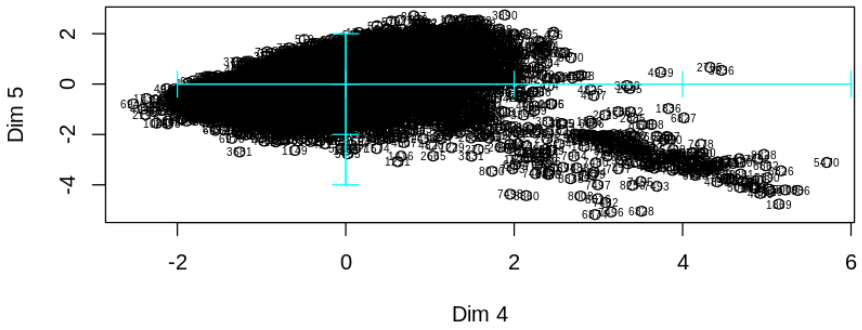
Subspace (X: 2 , Y: 5)



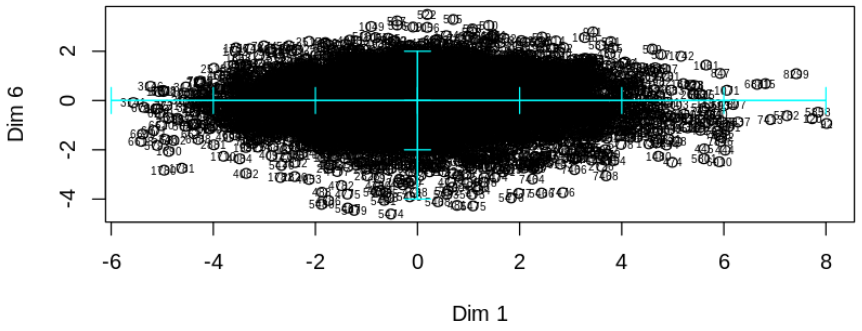
Subspace (X: 3 , Y: 5)



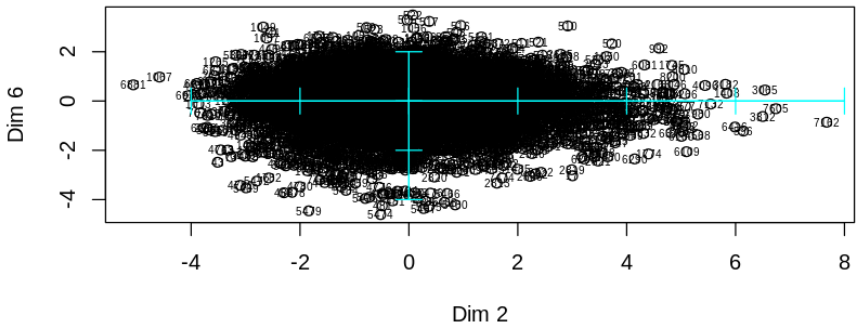
Subspace (X: 4 , Y: 5)



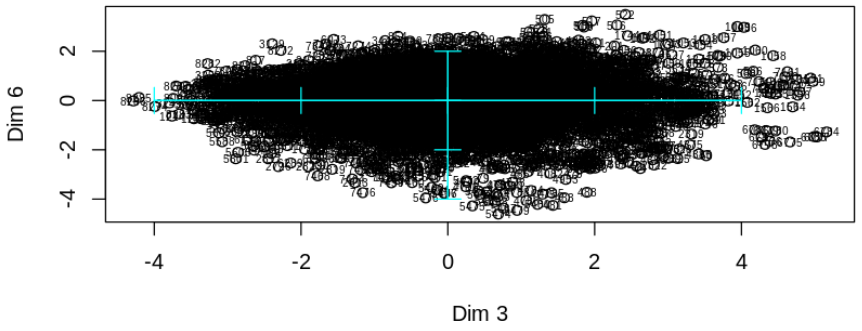
Subspace (X: 1 , Y: 6)



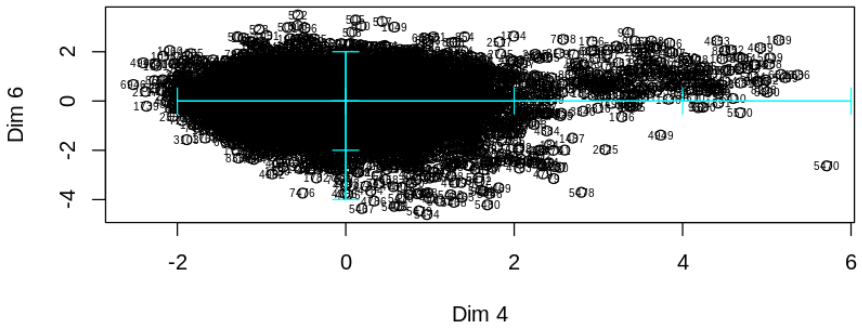
Subspace (X: 2 , Y: 6)



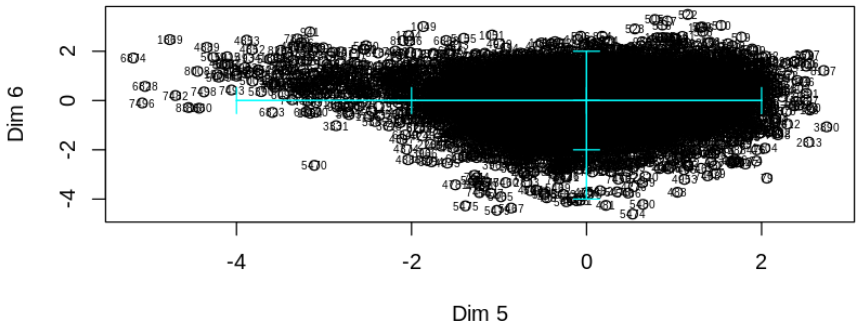
Subspace (X: 3 , Y: 6)



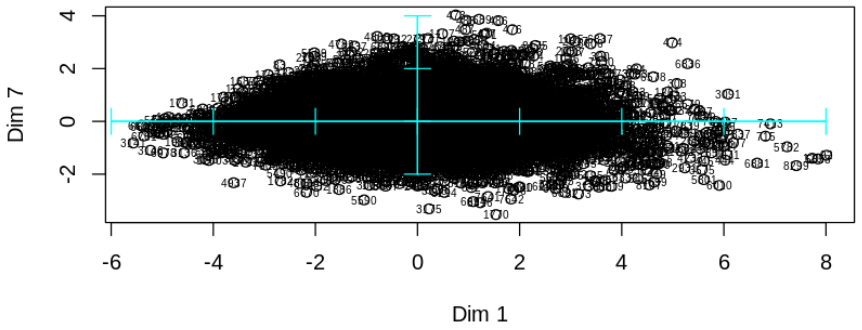
Subspace (X: 4 , Y: 6)



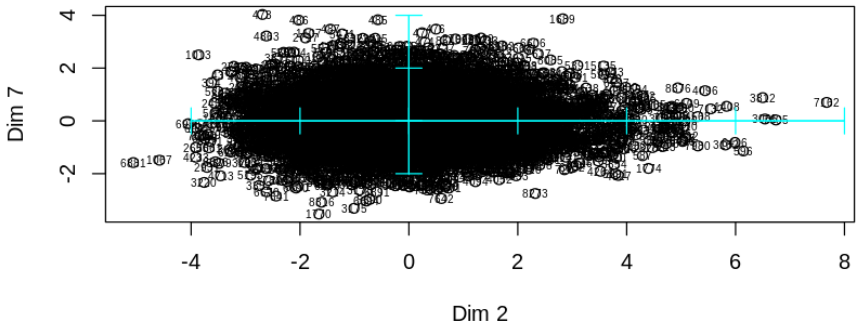
Subspace (X: 5 , Y: 6)



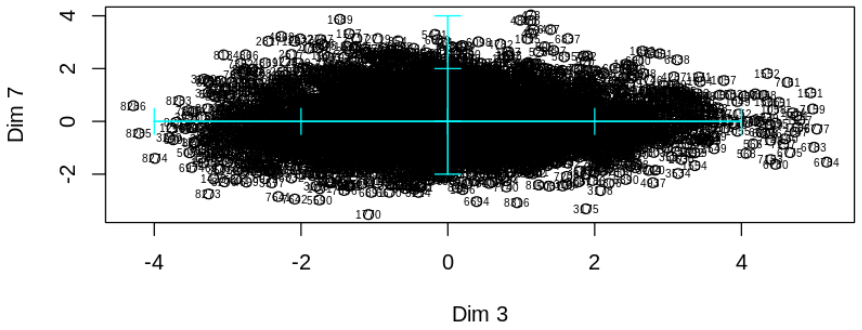
Subspace (X: 1 , Y: 7)



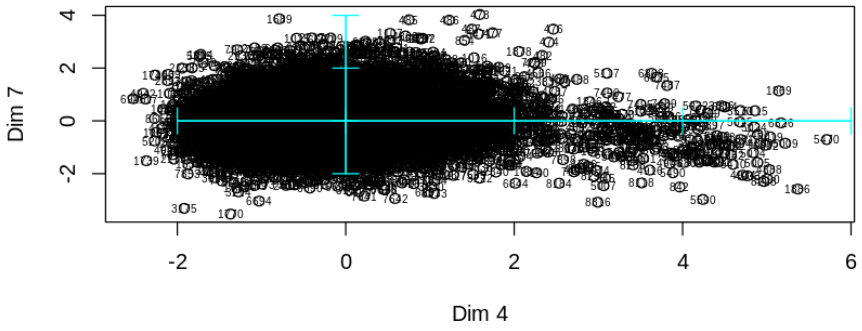
Subspace (X: 2 , Y: 7)



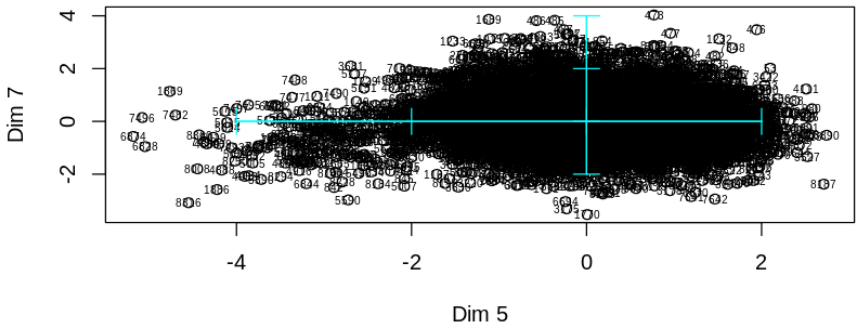
Subspace (X: 3 , Y: 7)



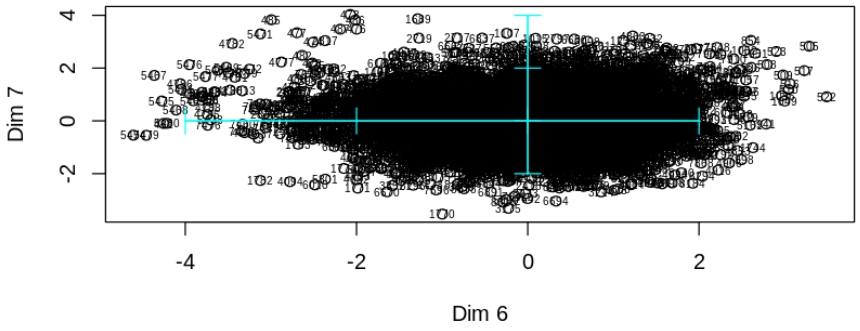
Subspace (X: 4 , Y: 7)



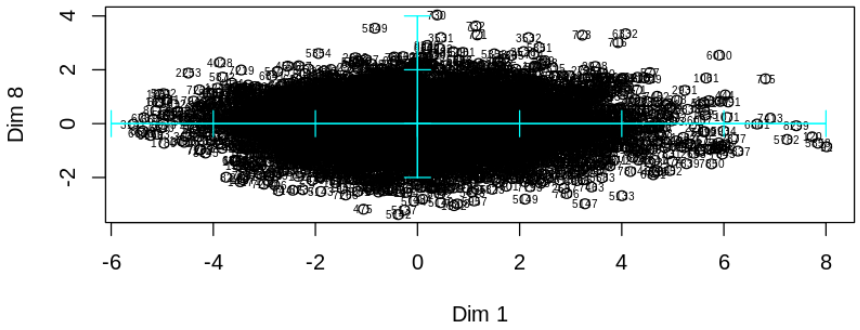
Subspace (X: 5 , Y: 7)



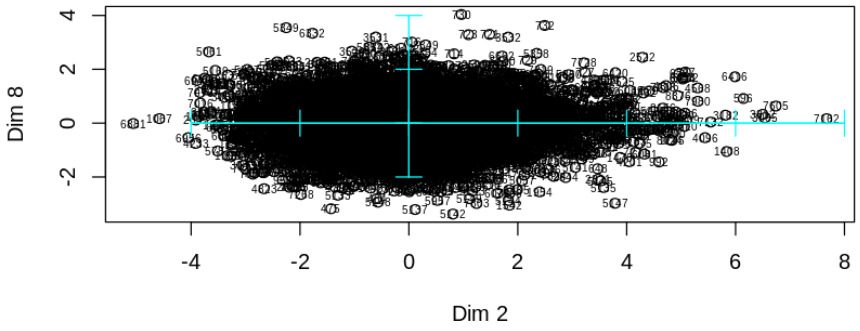
Subspace (X: 6 , Y: 7)



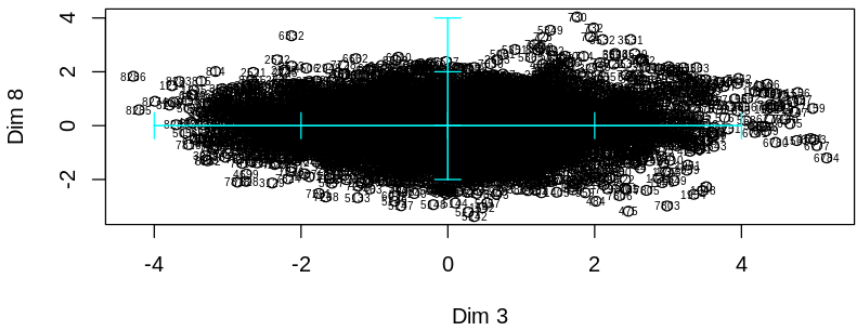
Subspace (X: 1 , Y: 8)



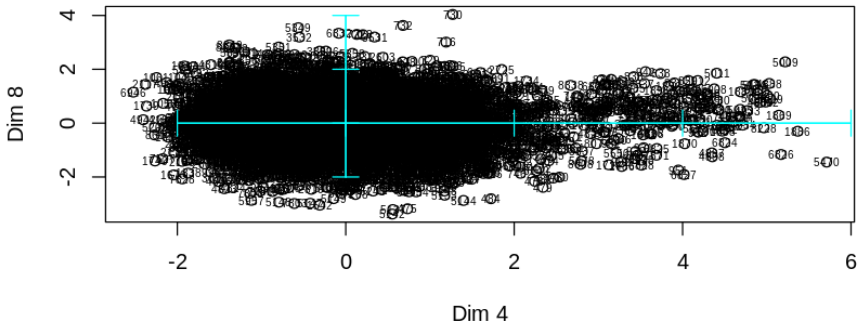
Subspace (X: 2 , Y: 8)



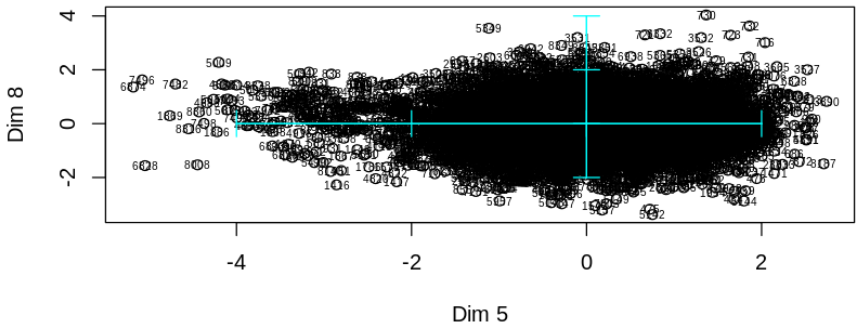
Subspace (X: 3 , Y: 8)



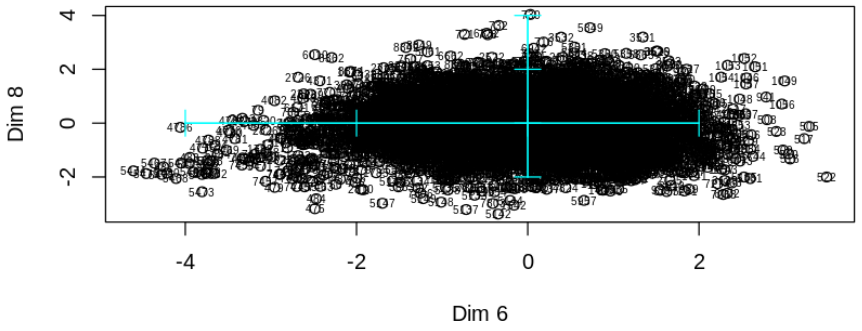
Subspace (X: 4 , Y: 8)



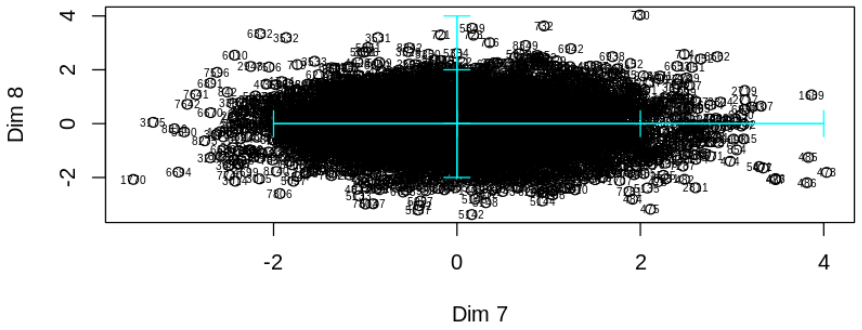
Subspace (X: 5 , Y: 8)



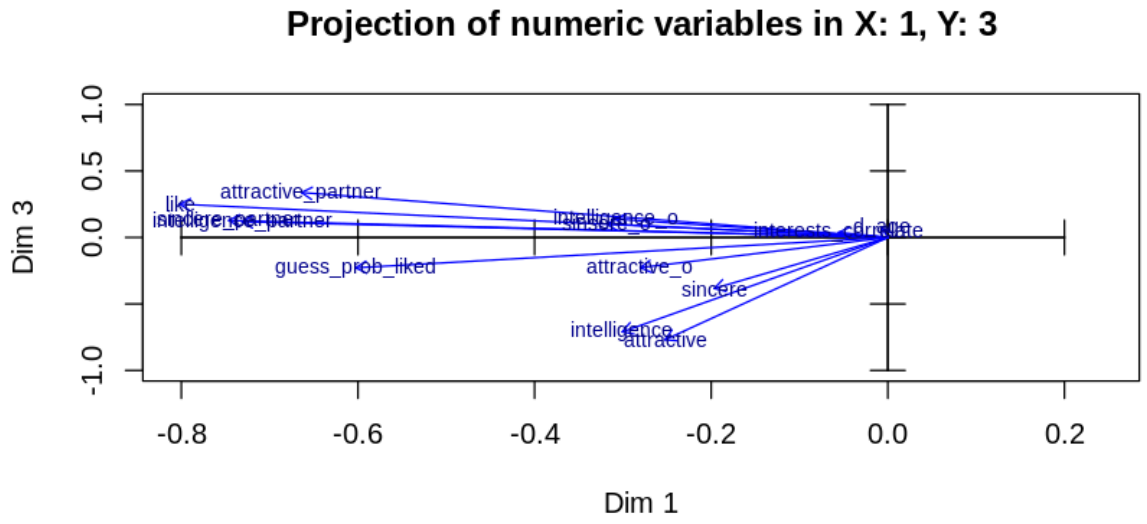
Subspace (X: 6 , Y: 8)

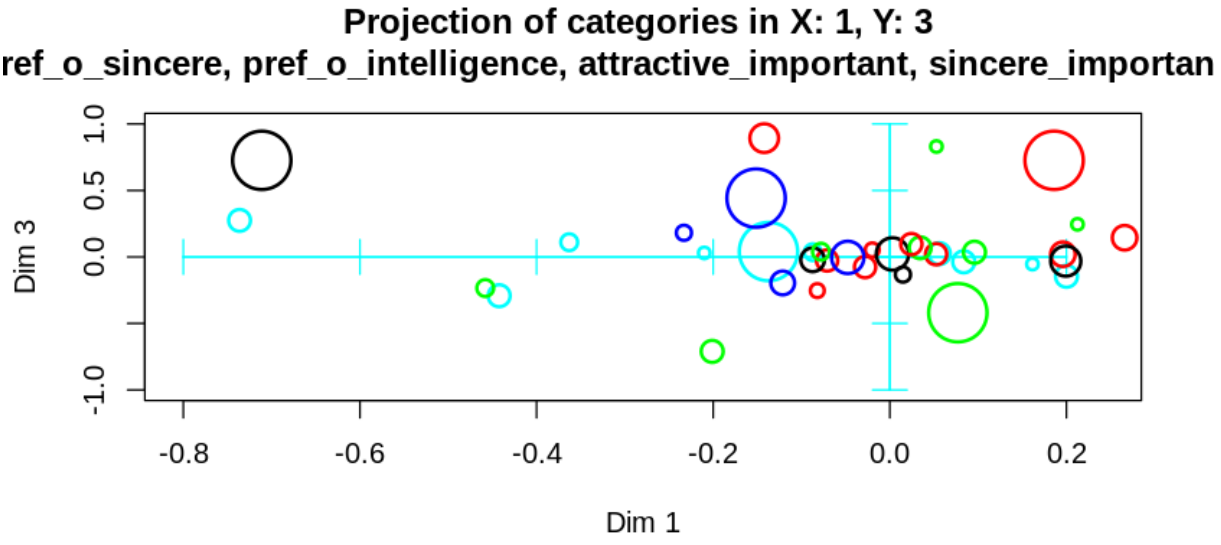


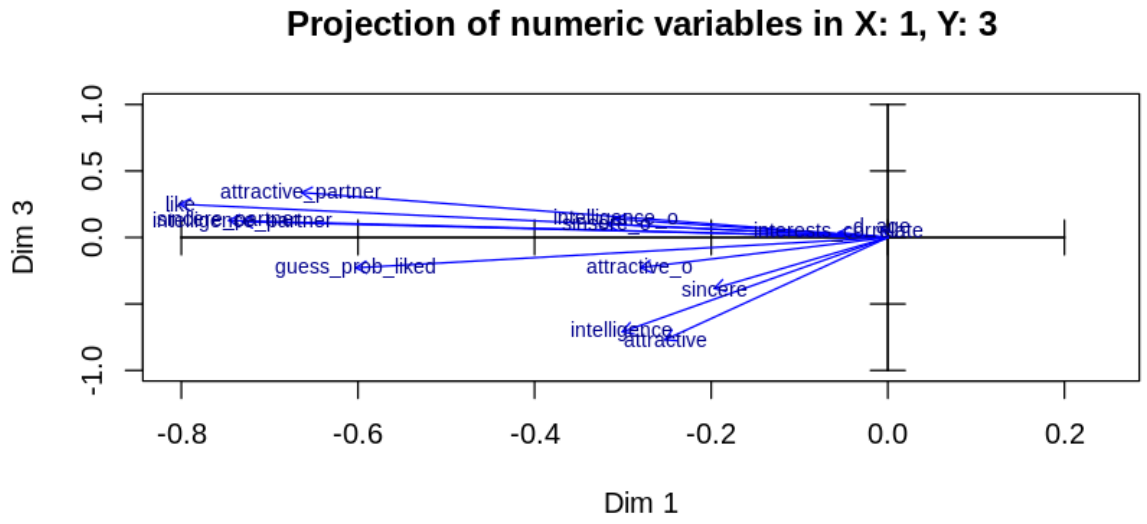
Subspace (X: 7 , Y: 8)

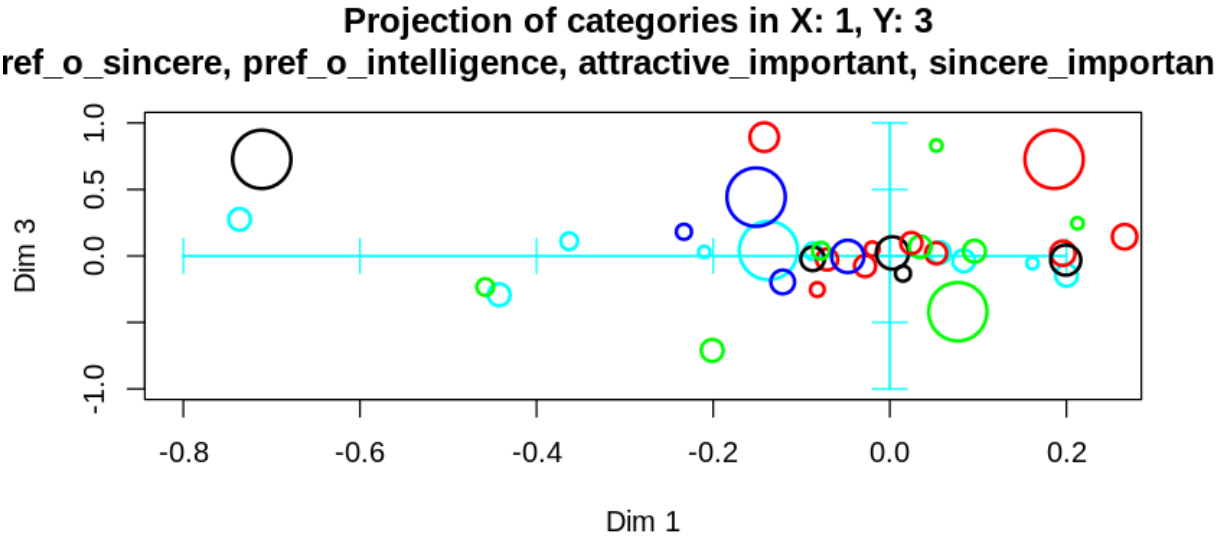


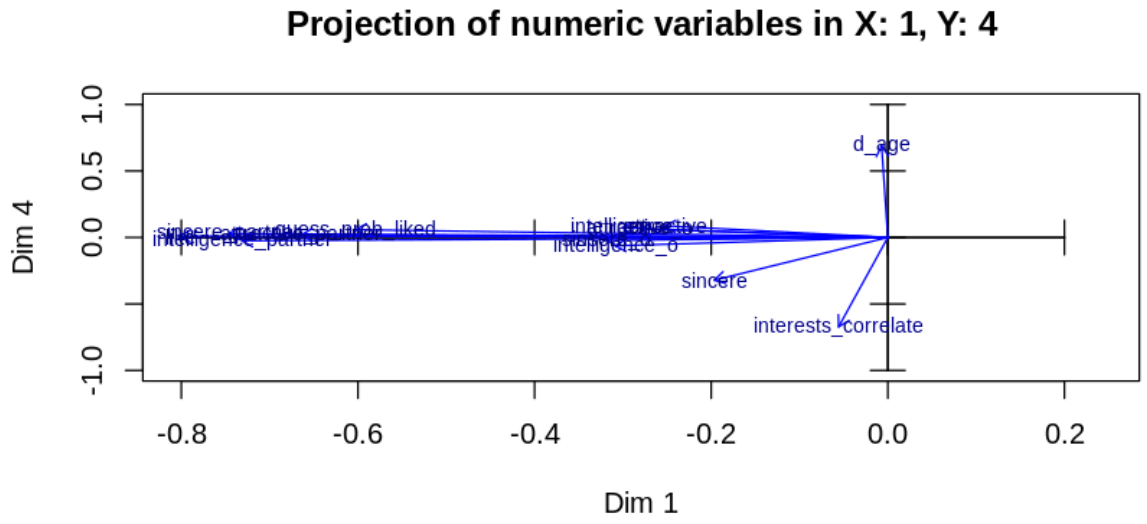
From all the plots obtained, we can clearly see that the subspaces (1,2), (1,3) ad (1,4) represent more variability than the others, now we are going to compare this 3 subspace in depth to decide which one are we working with.
We look at our possibilities:

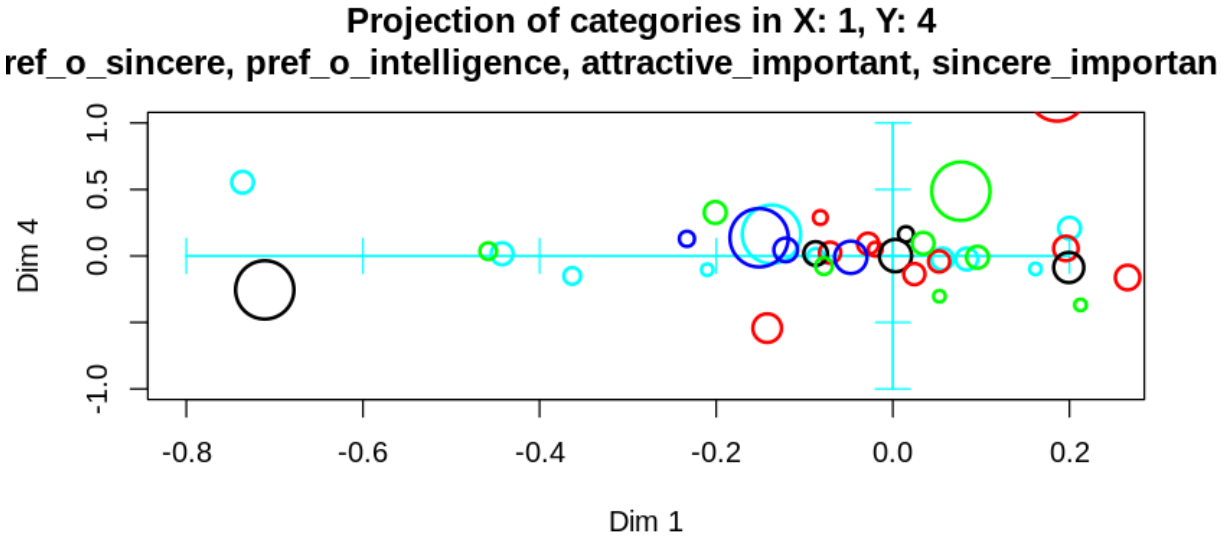





Projection of numeric variables in X: 1, Y: 3
Dim 3
1.0
0.5
0.0
-1.0
attractive_partner
guess_prob_liked
attractive_o
sincere
intelligence
attractive
-0.8
-0.6
-0.4
-0.2
0.0
0.2
Dim 1

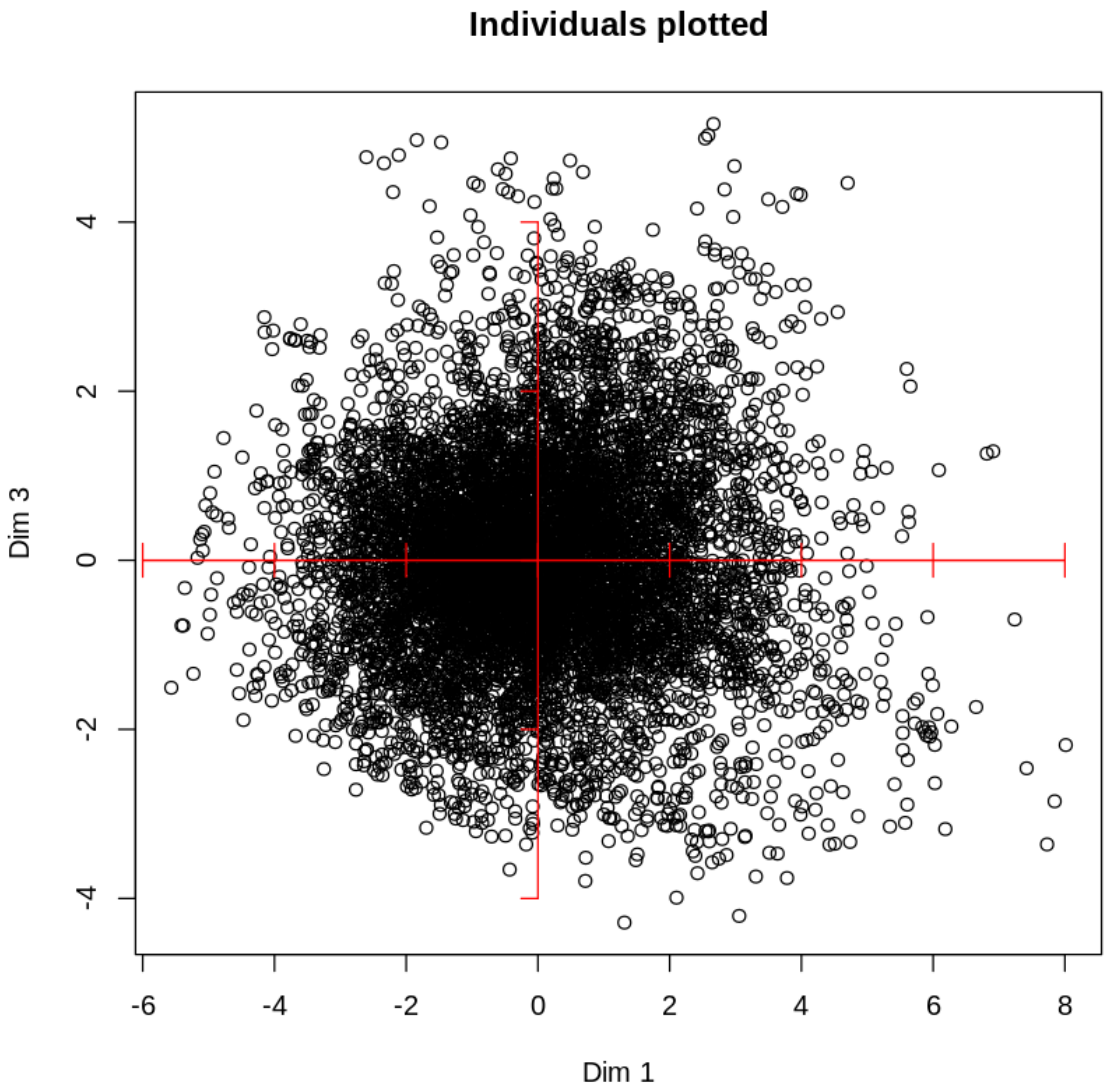

Projection of categories in X: 1, Y: 3
ref_o_sincere, pref_o_intelligence, attractive_important, sincere_importan
Dim 3
1.0
0.5
0.0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
Dim 1

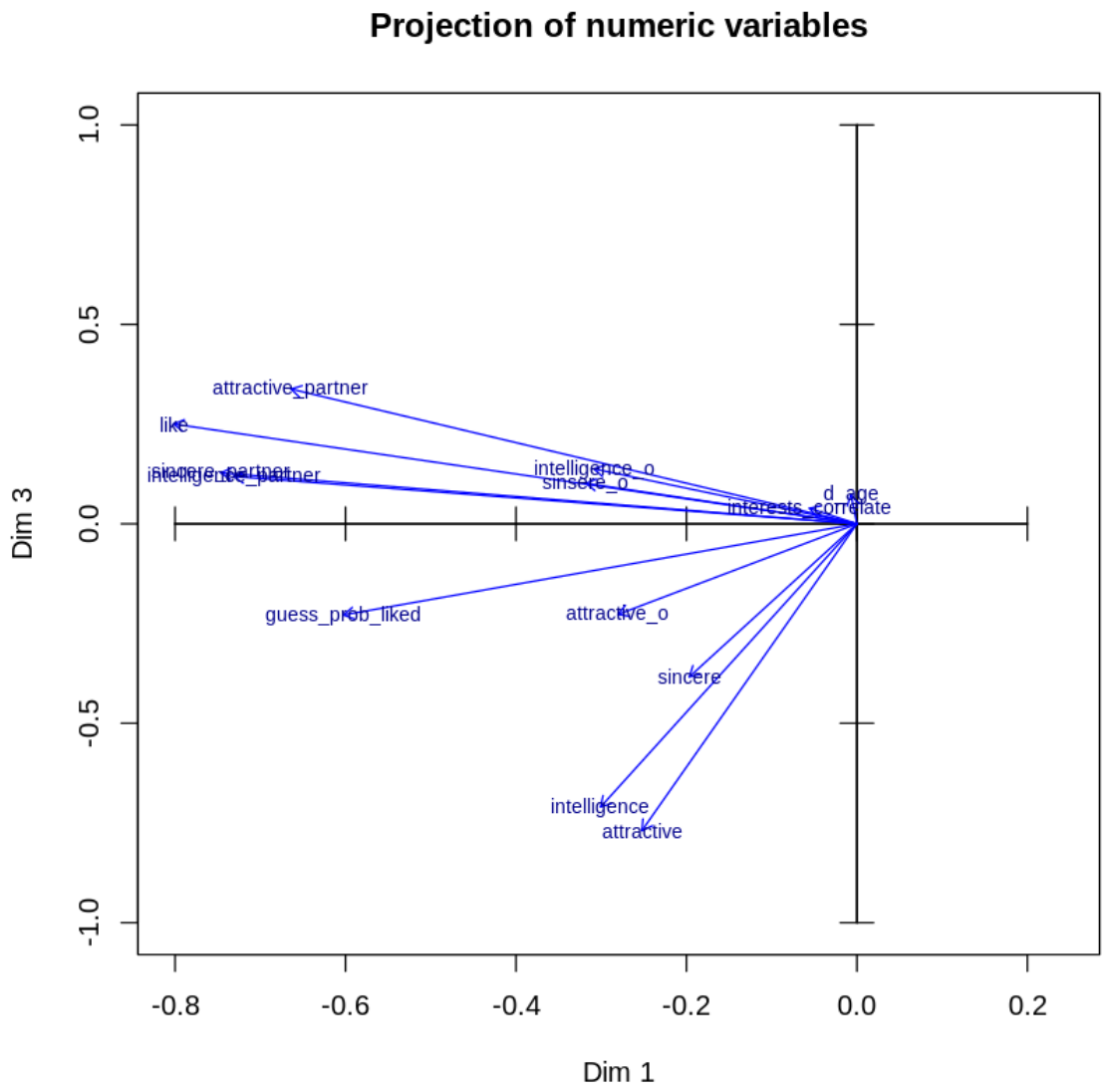




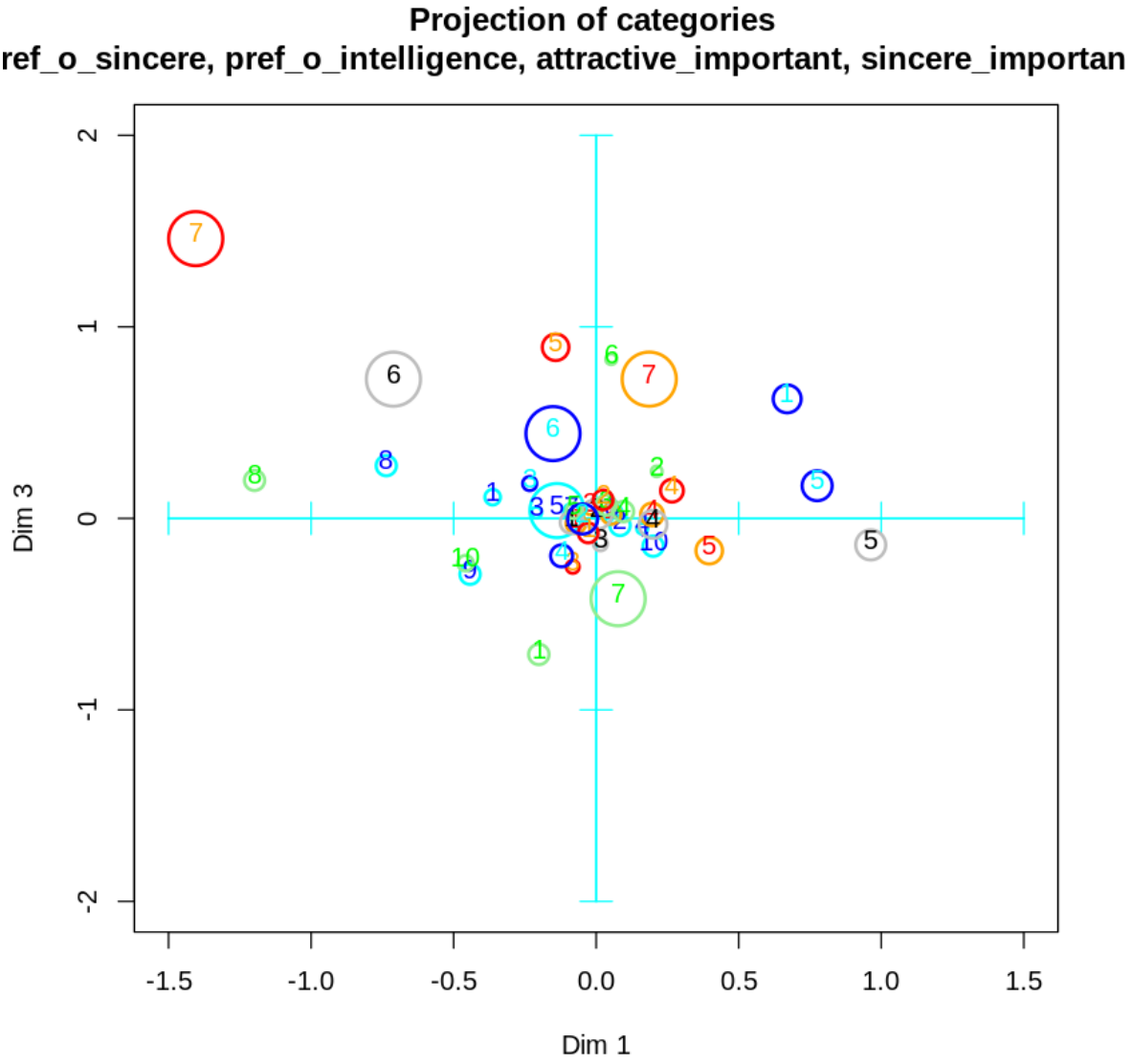
With the obtained results, we decided to use the subspace formed by the axis 1 and 3, because the vectors are more apart from each other. This means that the axis 3 represents a higher amount of variability with respect from the others, with a similar contribution value.

We plot the chosen subspace:





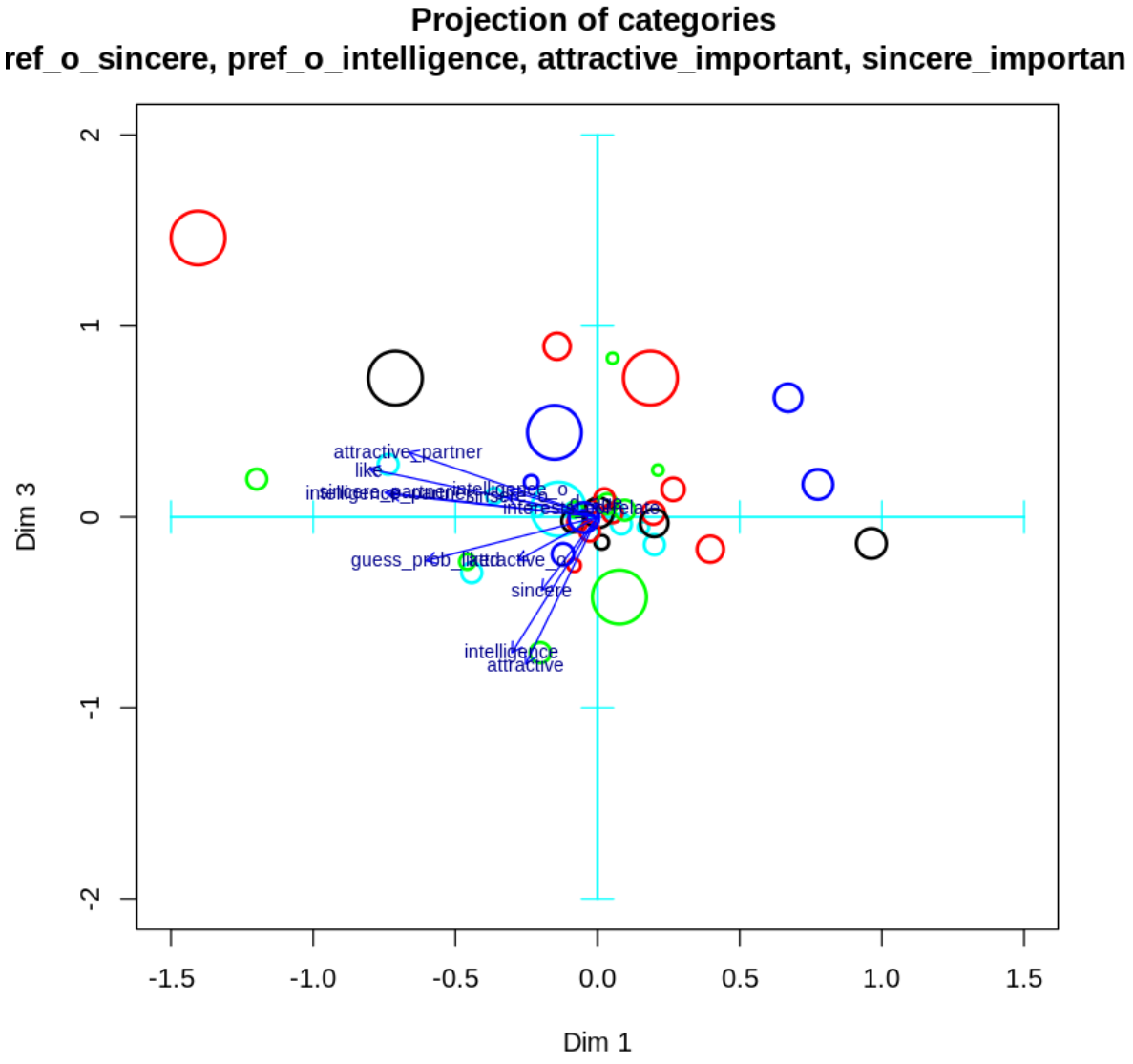
Dimension 1 shows how the partners have rated each individual and whether they have liked each other. On the other hand, dimension 3 shows the ratings that each individual gives himself or herself.



In the graph, we can identify the different categorical variables by means of the following caption:

- pref_o_attractive → Cyan
- pref_o_sincere → Orange
- pref_o_intelligence → Grey
- attractive_important → Green
- sincere_important → Red
- intelligence_important → Blue

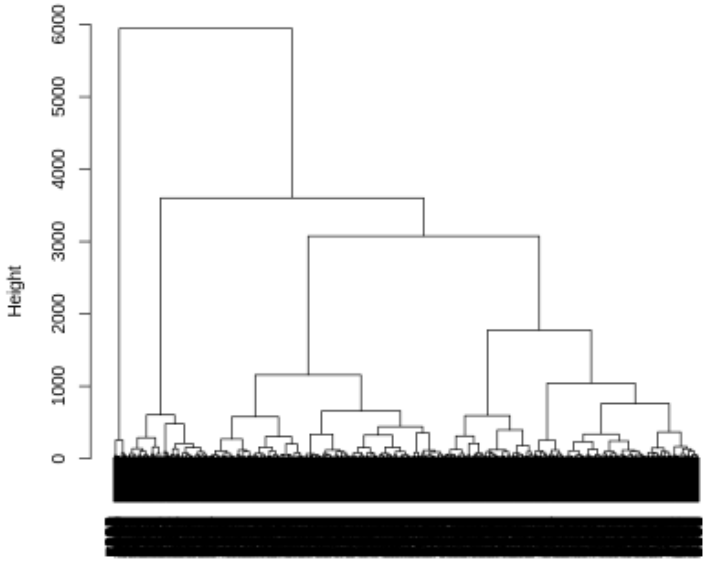
As we can see, there is no clear class separability through our categorical variables. Still, we can identify that high values related to physical attractiveness accumulate in the left sector of our graph. On the other hand, we see that high values related to intelligence tend to accumulate more towards the top. In this way we can conclude that more superficial people accumulate to the left, and people who are more focused on personality accumulate upwards, making the upper left sector more demanding people.


Projection of categories
ref_o_sincere, pref_o_intelligence, attractive_important, sincere_importan
attractive_partner
like
sincere
intelligence
attractive
Dim 3
Dim 1
2
1
0
-1
-2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Clustering

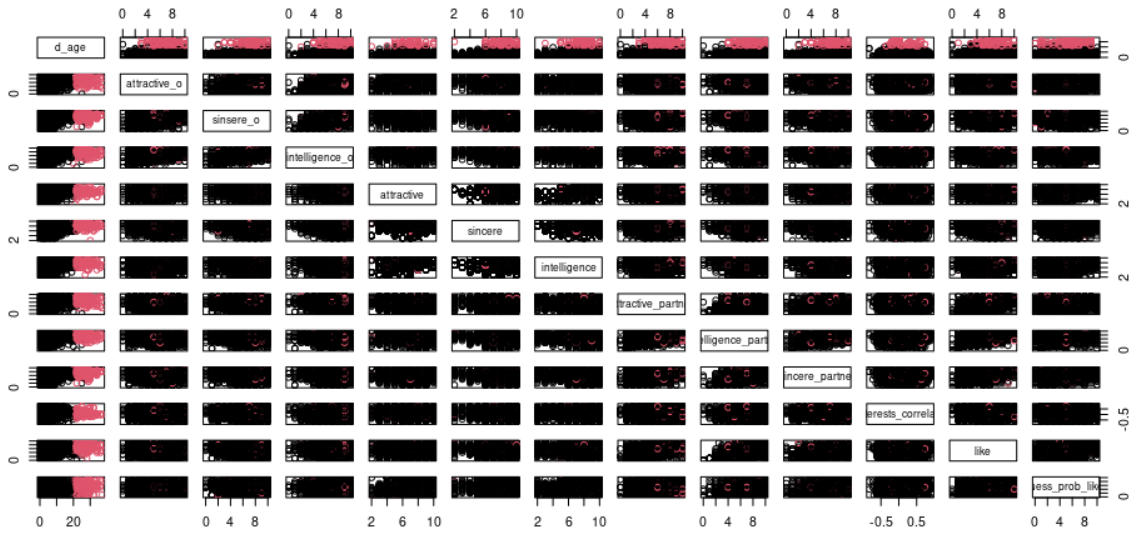
The clustering is used when we want to group individuals with similar variables. We only are going to work with numerical variables because to do the distance, we did the default metric which is Euclidean metric and it doesn't accept categorical variables.

To know how many clusters we have, we build the following dendrogram using the euclidean distance:



As we can see, the highest distance is between 3750 and 6000 approximately. It implies that we have to cut with 2 clusters.

To know more about the partition, we did the following pairplot of the cleaned data to know more about the clusters:



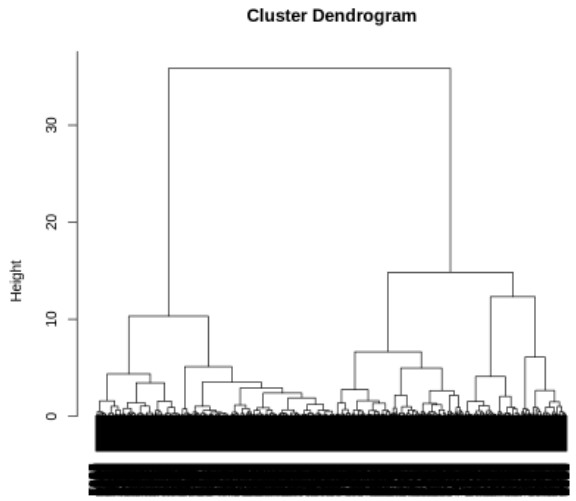
As we can see, the best variable that will help us to separate correctly the data in the 2 clusters is "d_age" because as we can see in the plots, when it is in the X axis, we can appreciate 2 sides. One side between 0 and 18 in black color, that means it is one of the clusters and the other side between 19 and 30 with red color, that means it is the other cluster. So the majority of the individuals of one cluster are between 0 and 18 and the majority of individuals of the other cluster are between 19 and 30.

Another thing that we can see is if we have a look on the plot with the variables "interest_correlated" and "d_age", we can see the separation that I mentioned wit the "d_age" and a pyramid, it means that when the difference of ages increases, the interest that they have in common decreases because they are more near to 0, specially when we see the pink cluster.

The last thing that I can see if we have a look on the plots with "d_age" and the other variables in the other axis, We can appreciate that the pink cluster usually have better punctuation in the variables that are ratings (those are all the numerical variables excepts "d_age", "interest_correlated" and "guess_prob_loked") than in the black cluster. I mean, in the pink cluster (with higher ages) the minimum punctuation of the majority of the pink cluster members are higher than the black cluster with less different ages. For example, if we have a look in "sinsere" and "d_age" variables, for the pink cluster with more different of ages between the partner, the punctuation for the sincerity of itself is between 6 to 10, but with more similar ages in the black cluster, the rating is between 2 to 10. There are 4 points of difference in the minimum punctuation of both clusters. This is the more extreme case, but with the other variables happens the same (the minimum rating is higher in the majority of the members in the pink cluster than in the majority of members in the black cluster).

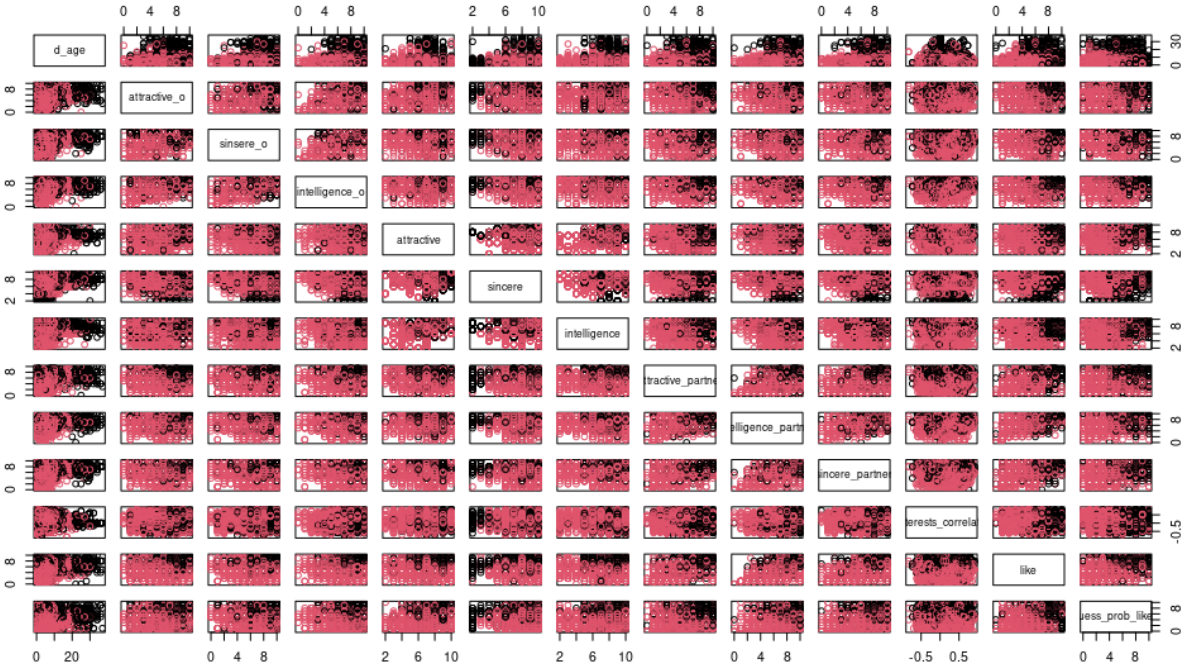
The observations are pretty well, but the main problem is that we only have 196 individuals in cluster 2 and more than 8000 in cluster 1.

Let’s try to do it with the distance matrix with gower distance but with only numerical variables, because we did it before with only numerical variables. First of all, we have to see how many clusters we have doing a dendrogram. We build the next one:



As we can see, to make a cut in the dendrogram, as in the other dendrogram, the larger distance is with 2 clusters. We are going to have 2 clusters for this reason.

And as before, to know more about the clusters, we made the following pairplot:



As we can see, it happens like in the other representation. The best variable to separate the data in two clusters is "d_age" because the pink and black clusters are more or less separated. The majority of the members in the pink cluster are on the left side and the majority of the members in the black cluster are in the middle-right side. But, this is worse separated than in the other, but we can see some patterns:

When "d_age" is in the X axis, we usually find the majority of pink clusters in the left side and the black individuals in the middle-right side. We can see it better when the Y axis is one of the next variables: "attractive_o","sinsere_o","intelligence_o" "intelligence_partner" and "sincere_partner". And with these variables, we can see, as the other representation, that the minimum value of the majority of the members in the black cluster is lower than the minimum value of the majority of the members in black cluster.

The last thing I'll mention is that we can appreciate a diagonal separation (top-left to bottom-right) if we look at the variables "guess_prob_likes" and "d_age". If we look down the diagonal, we can see the majority of individuals in pink cluster and if we look above the diagonal, we can see the majority of the individuals in black cluster. This diagonal means for the members of the pink cluster that when the difference of ages increases, the rating about what do you think about the other person liked you decreases and for the members in the black cluster: when the difference of ages decreases, the rating about what do you think about the other person liked you, increases.

This partition, doing the distance matrix, the distribution of the individuals in each cluster is more balanced (there are approximately 4000 individuals in each cluster). And we have 2 clusters too. For those two reasons, we will do the study with the distance matrix and 2 clusters

Profiling

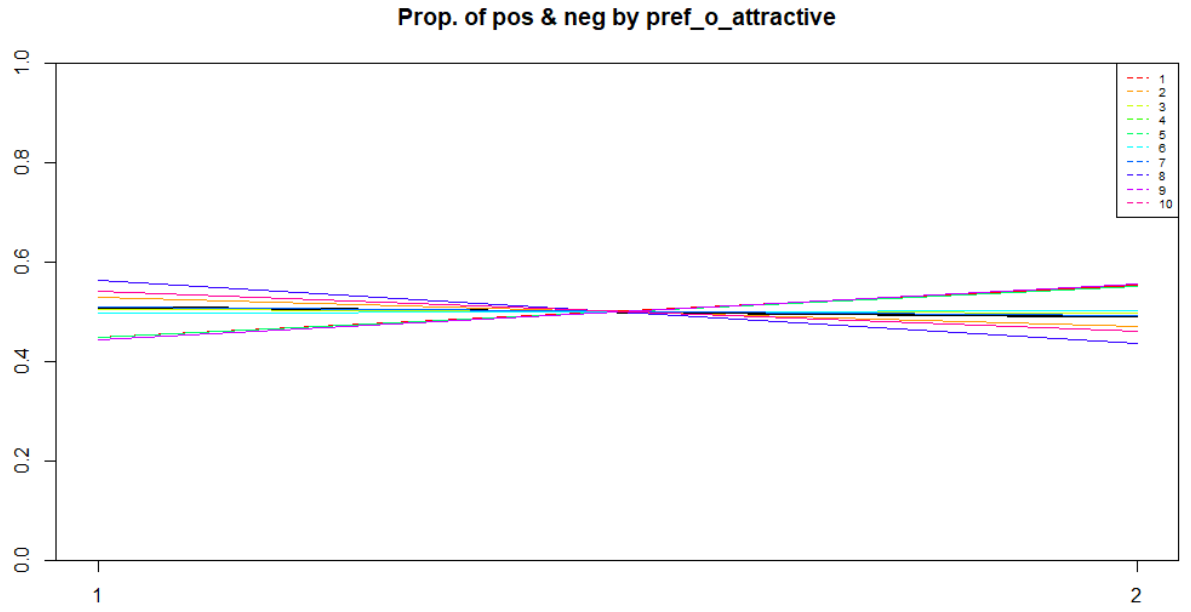
Data profiling, also known as profiling, is a statistical technique that involves examining the information contained within a data source. It involves analyzing the relationships between the data, its structure, and the information provided by each variable. Prior to conducting the profiling analysis, the data set was subjected to clustering to group individuals into different clusters. As a result, a new variable called “cluster” was introduced, which indicates the group to which each individual belongs. The data set is now divided into two distinct clusters.

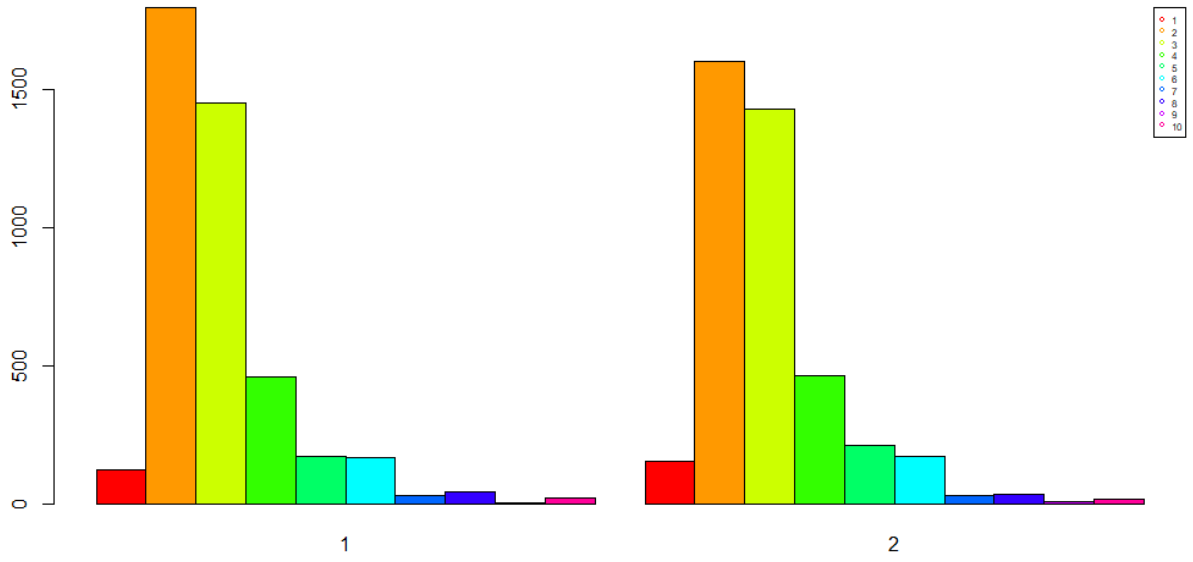
d_age

The first variable we are going to observe and analyze its relationship with the clusters is d_age, that represents the difference of ages between partners.
In those two plots we can observe that the median and the two quartiles are exactly the same, but the mean of the cluster 1 is higher because the individuals with the most value in this variable belongs to that cluster.

pref_o_attractive

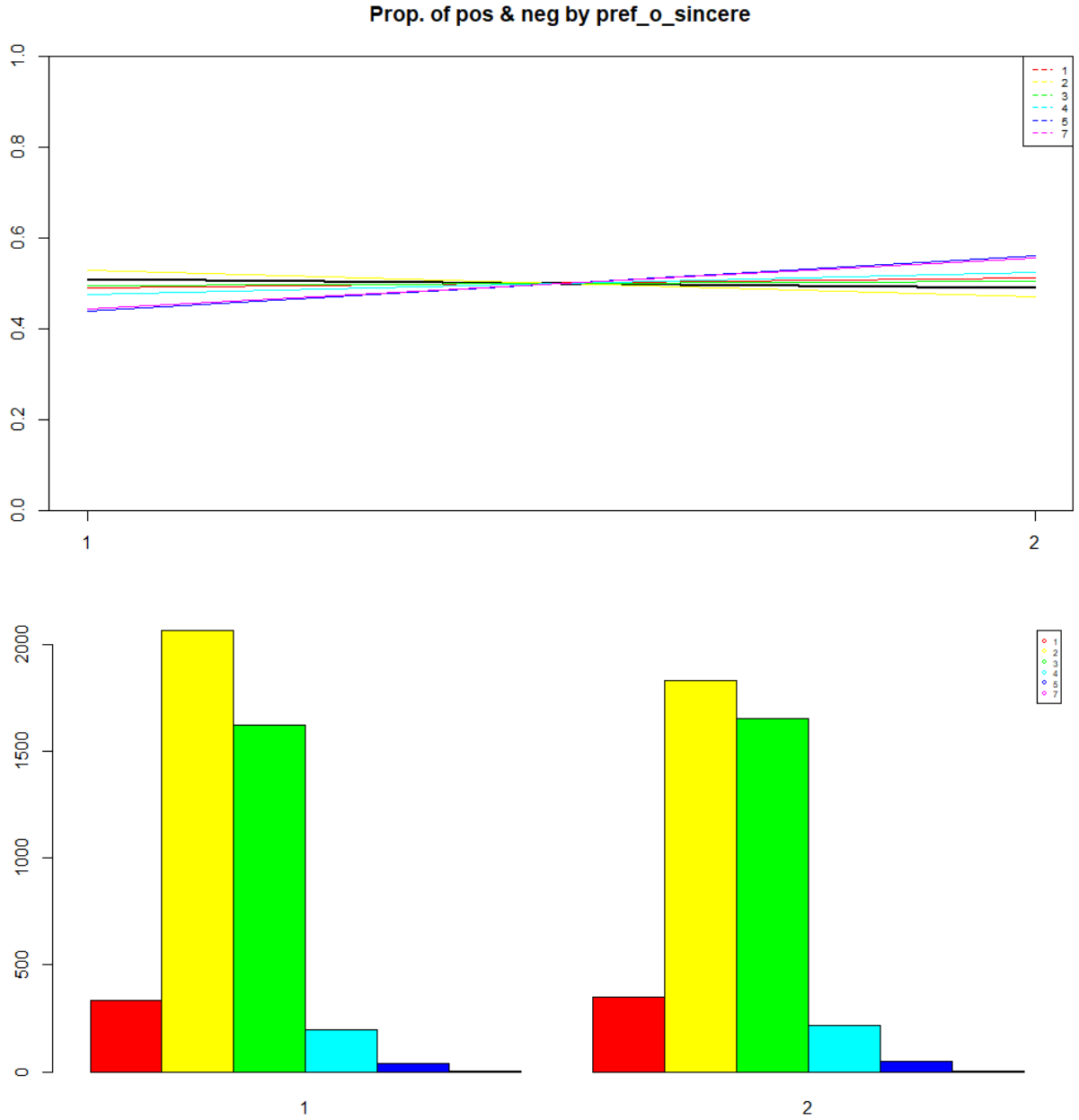
The next variable we are going to analyze is the relationship between the cluster and variable pref_o_attractive, which represents the rating about how important the partner think is attractiveness.
We can observe that there's no something differential between the clusters. They both are equally distributed.





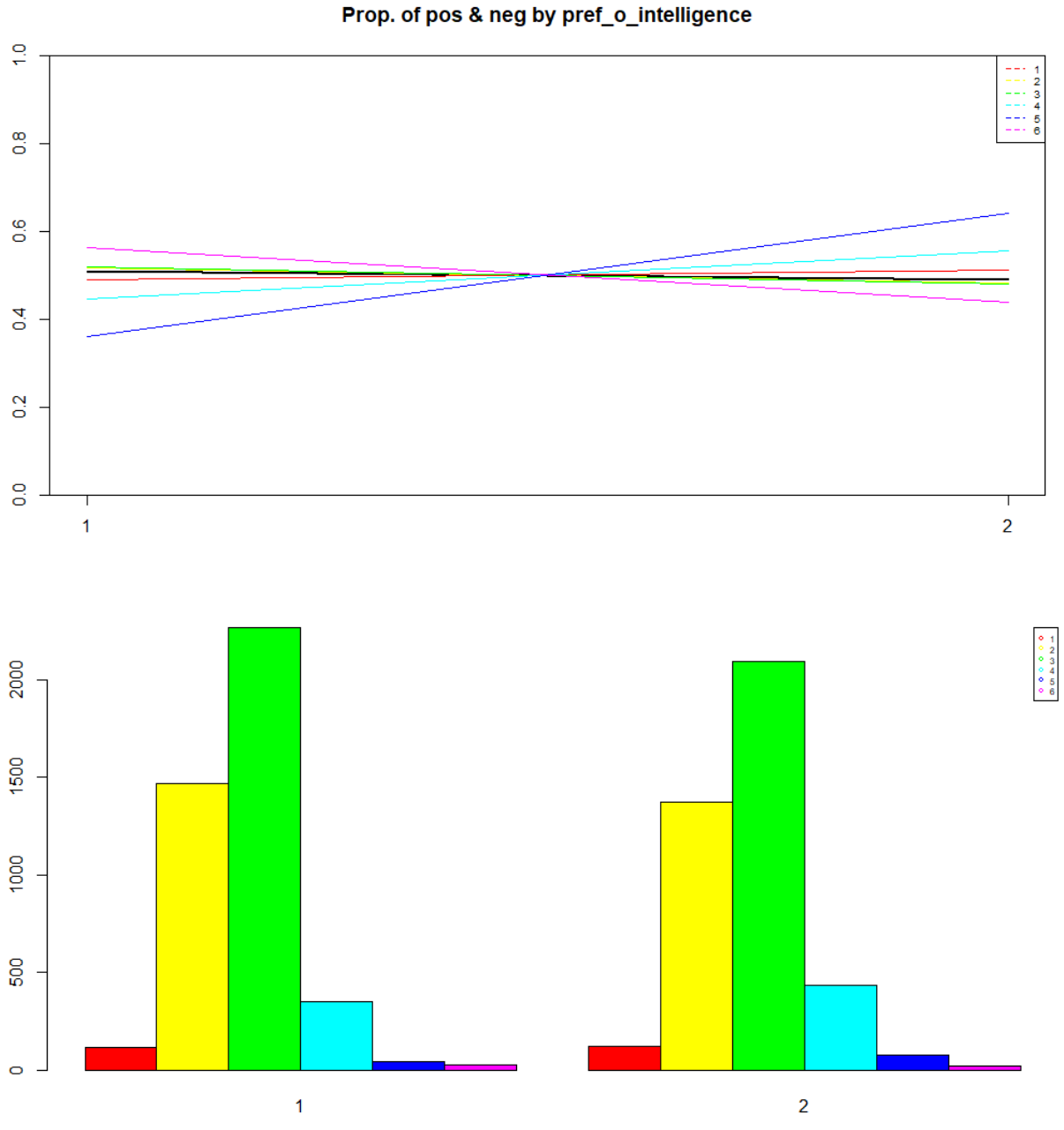
pref_o_sincere

The next variable we are going to analyze is the relationship between the cluster and variable pref_o_sincere, which represents the rating about how important the partner think is sincerity.
We can observe that there's no something differential between the clusters. They both are equally distributed.



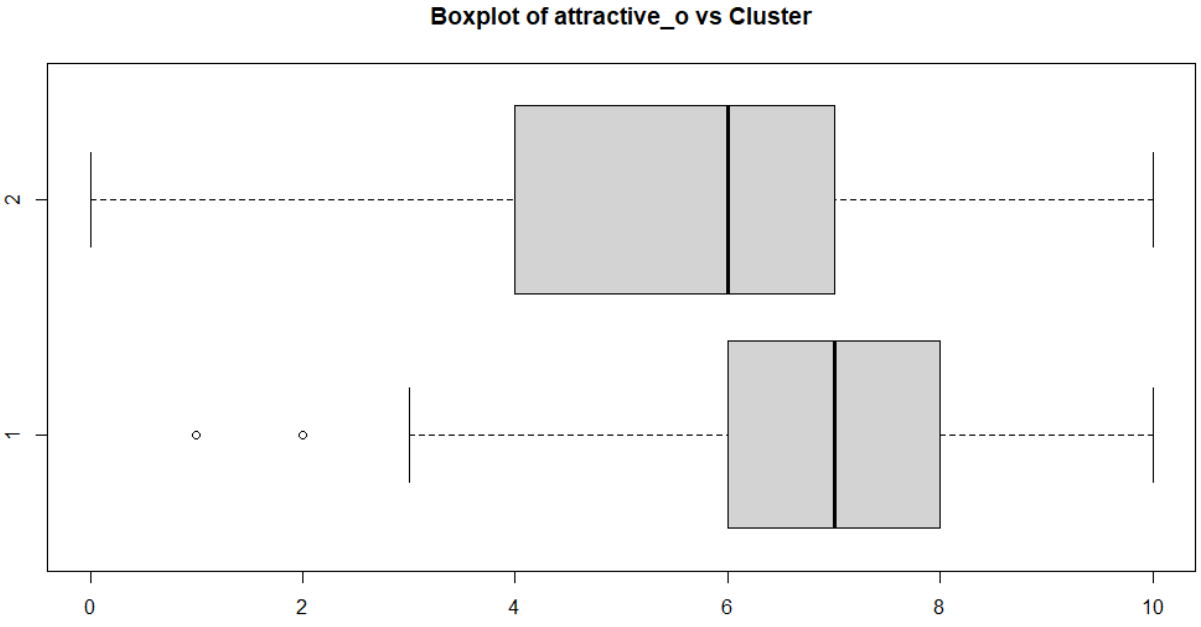
pref_o_intelligence

The next variable we are going to analyze is the relationship between the cluster and variable pref_o_intelligence, which represents the rating about how important the partner think is intelligence.
We can observe that there's no something differential between the clusters. They both are almost equally distributed, because we can see that the probability of the factor 5 being in the cluster two is relatively higher.



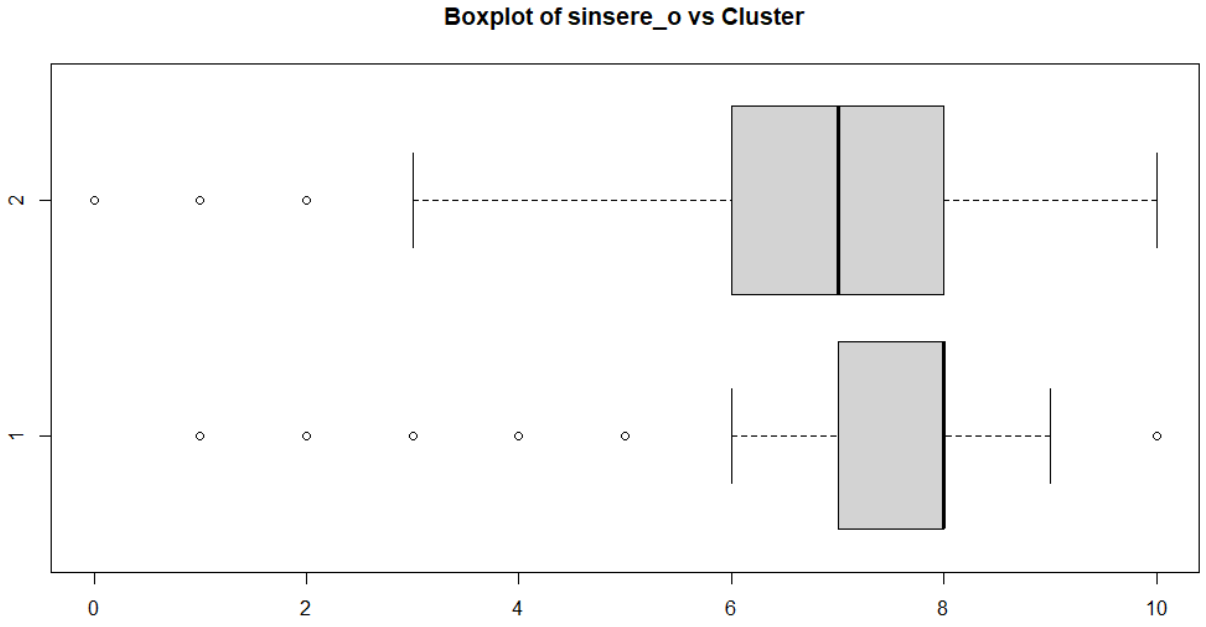
attractive_o

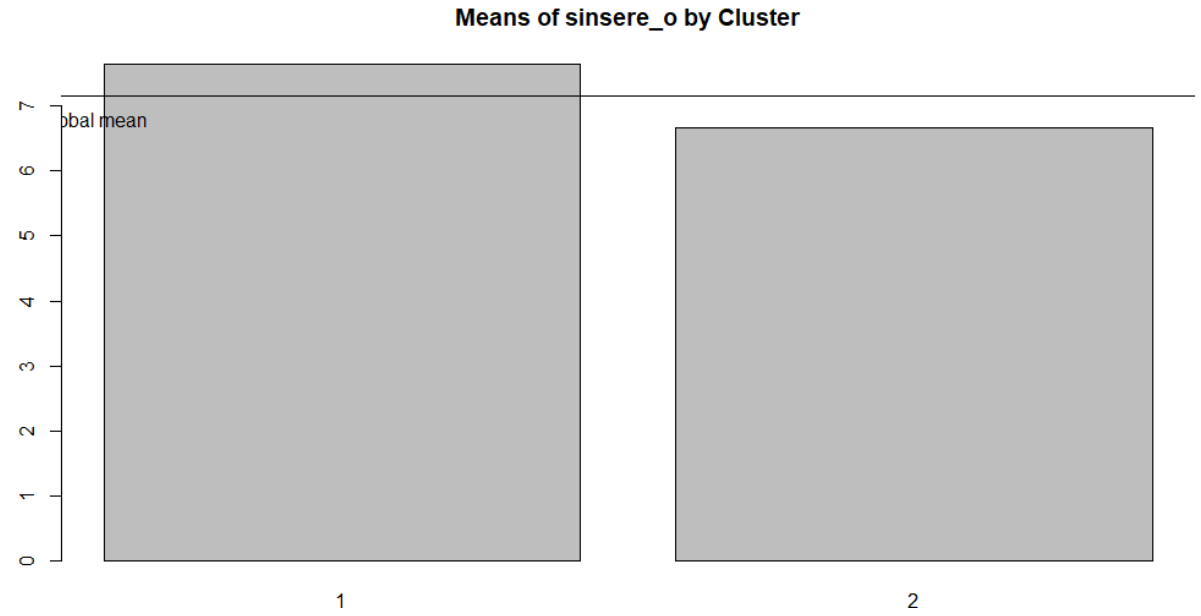
The next variable we are going to analyze is the relationship between the cluster and variable attractive_o, which represents the rating about attractiveness made by the partner at night of the event.
We can observe that the individuals in the first cluster are more likely to get higher ratings, between 6 to 8. The individuals in the second cluster have ratings more distributed, with the big amount between 4 to 7.
The difference between the means is 0.985. It's almost 1 point better ranked.
We could conclude that the people thinks that the individuals in cluster 1 are more attractive.



sinsere_o

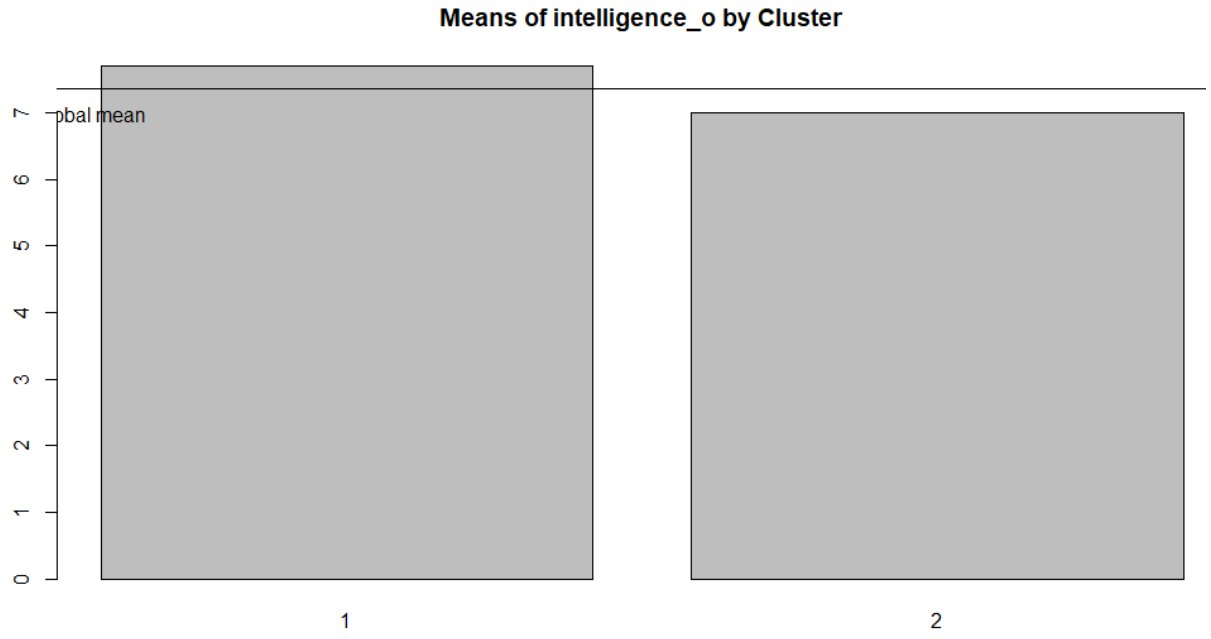
The next variable we are going to analyze is the relationship between the cluster and variable sinsere_o, which represents the rating about sincerity made by the partner at night of the event.
We can observe that the individuals in the first cluster are more likely to get higher ratings, between 7 to 8. The individuals in the second cluster have ratings a little bit more distributed, with the big amount between 6 to 8.
The difference between the means is 0.973.
We could conclude that the people thinks that the individuals in cluster 1 are more sincere.





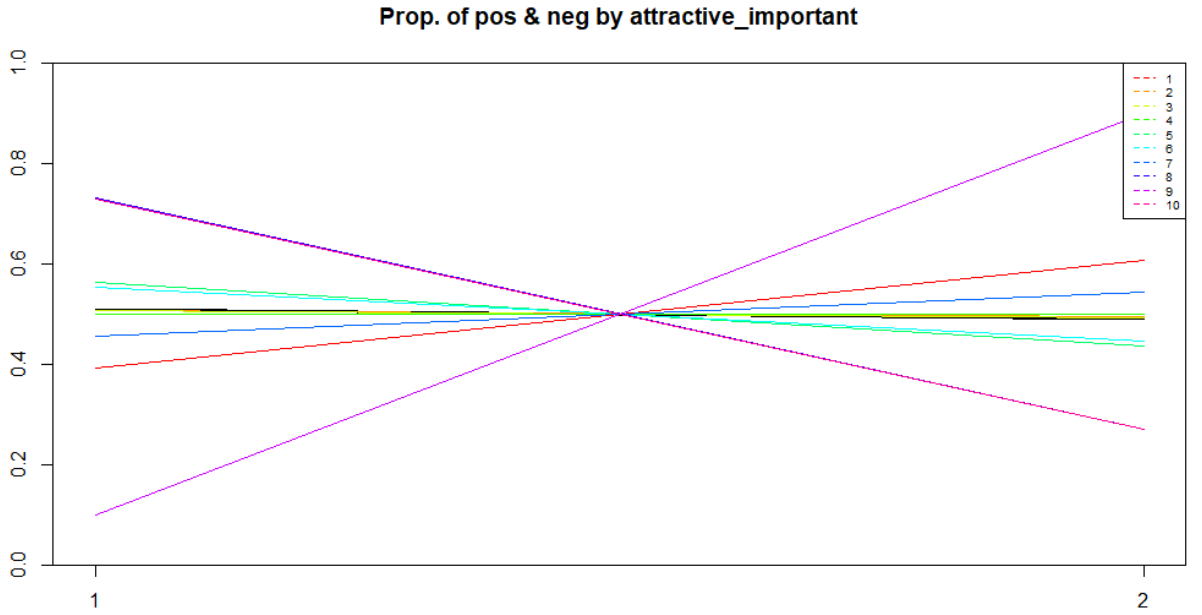
intelligence_o

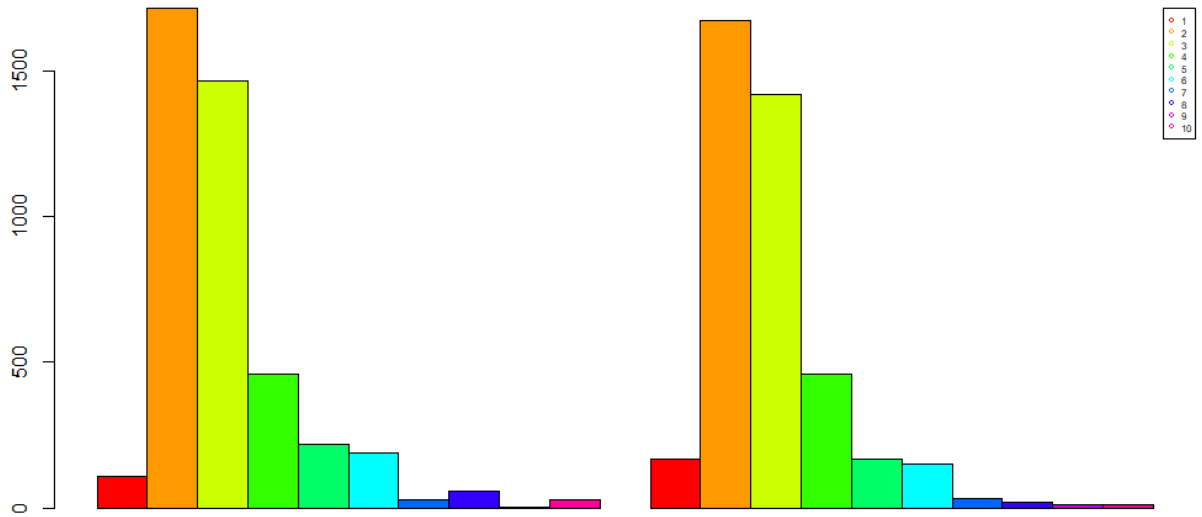
The next variable we are going to analyze is the relationship between the cluster and variable intelligence_o, which represents the rating about intelligence made by the partner at night of the event.
We can observe that the individuals in the first cluster are more likely to get higher ratings, between 7 to 9. The individuals in the second cluster have ratings a little bit more distributed, with the big amount between 6 to 8.
The difference between the means is 0.693, that is the lowest difference in relation to the 3 attributes we are analyzing.
We could conclude, but with less certainty, that the people thinks that the individuals in cluster 1 are more intelligent.



attractive_important

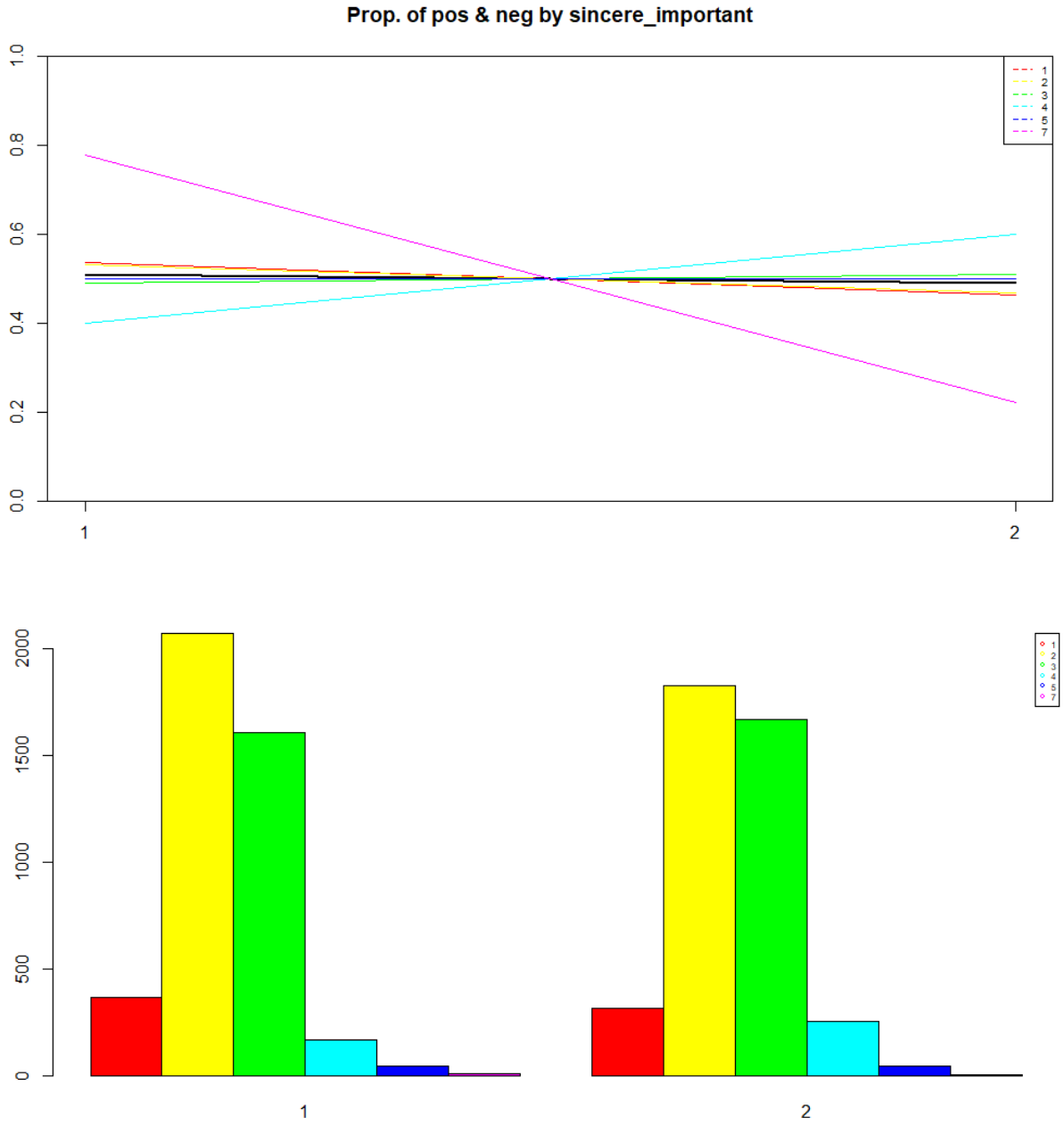
The next variable we are going to analyze is the relationship between the cluster and variable attractive_important, which represents the rating about how important the participant think is attractiveness.
We can observe that there's no something differential between the clusters in terms of the quantity of individuals, but we can see some visible slopes. That's because the factors with low individuals, although the difference in quantity is low, proportionally is high.





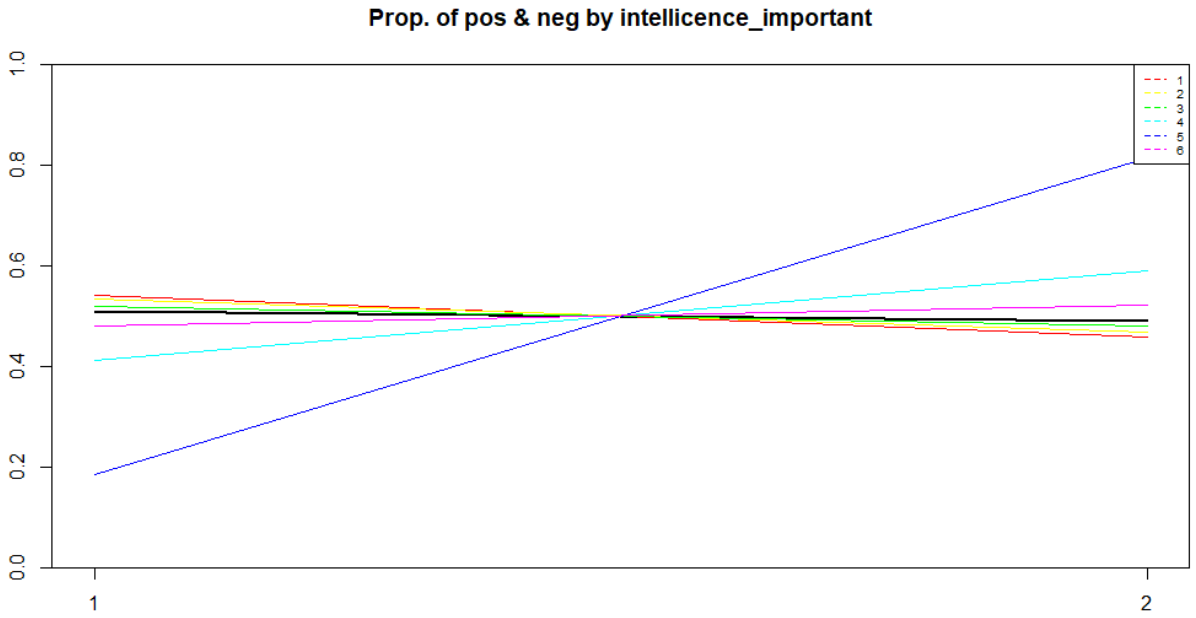
sincere_important

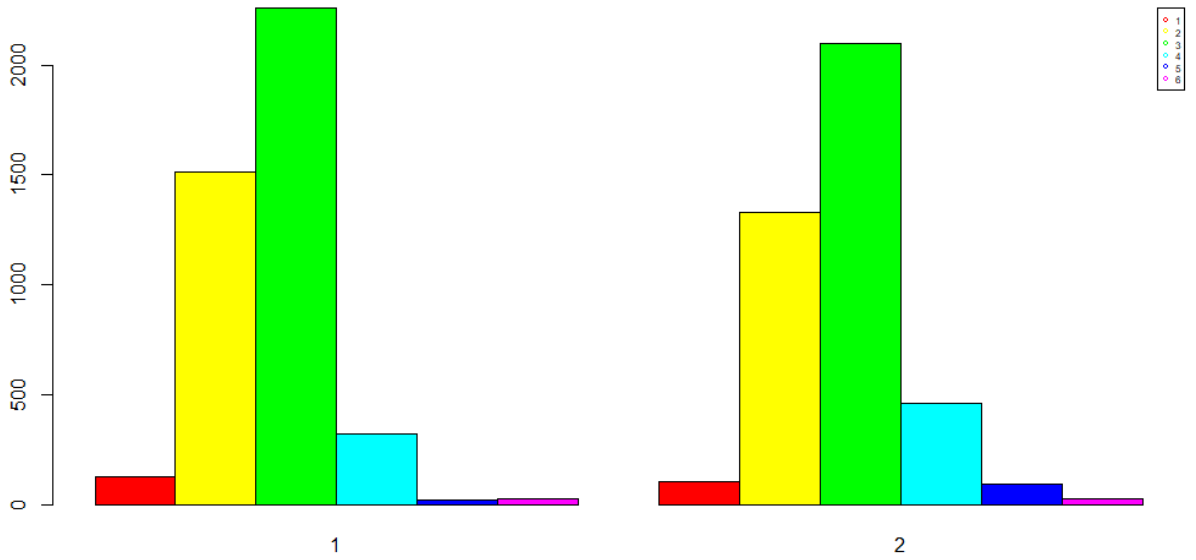
The next variable we are going to analyze is the relationship between the cluster and variable sincere_important, which represents the rating about how important the participant think is sincerity.
We can observe that there's no something differential between the clusters in terms of the quantity of individuals, but we can see some visible slopes. That's because the factors with low individuals, although the difference in quantity is low, proportionally is high.



intelligence_important

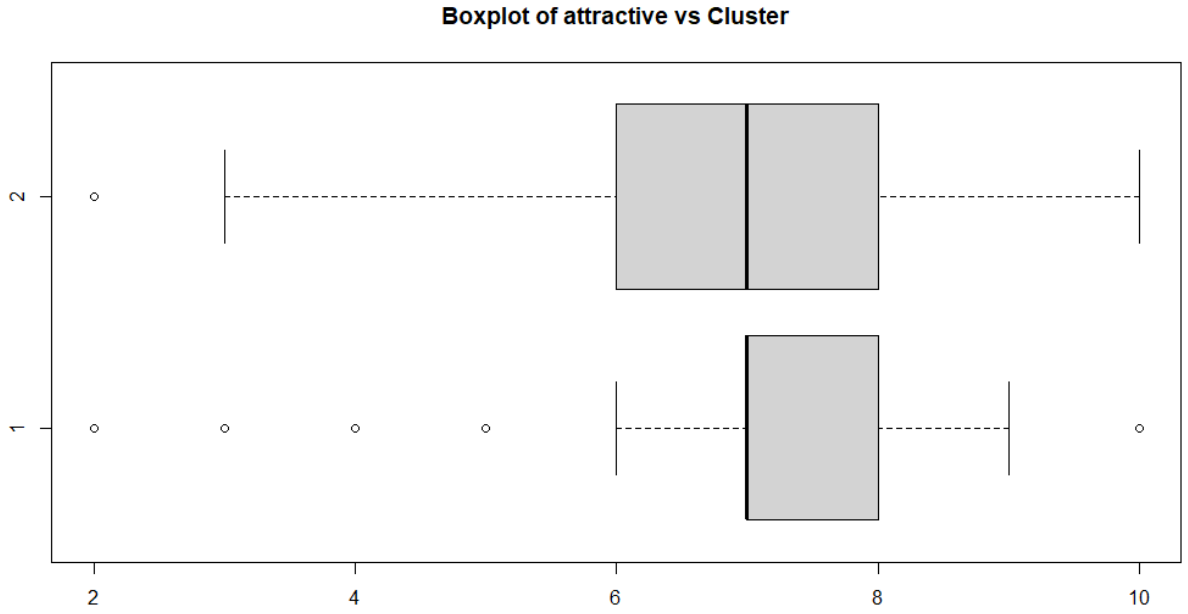
The next variable we are going to analyze is the relationship between the cluster and variable intelligence_important, which represents the rating about how important the participant think is intelligence.
We can observe that there's no something differential between the clusters in terms of the quantity of individuals, but we can see some visible slopes. That's because the factors with low individuals, although the difference in quantity is low, proportionally is high.





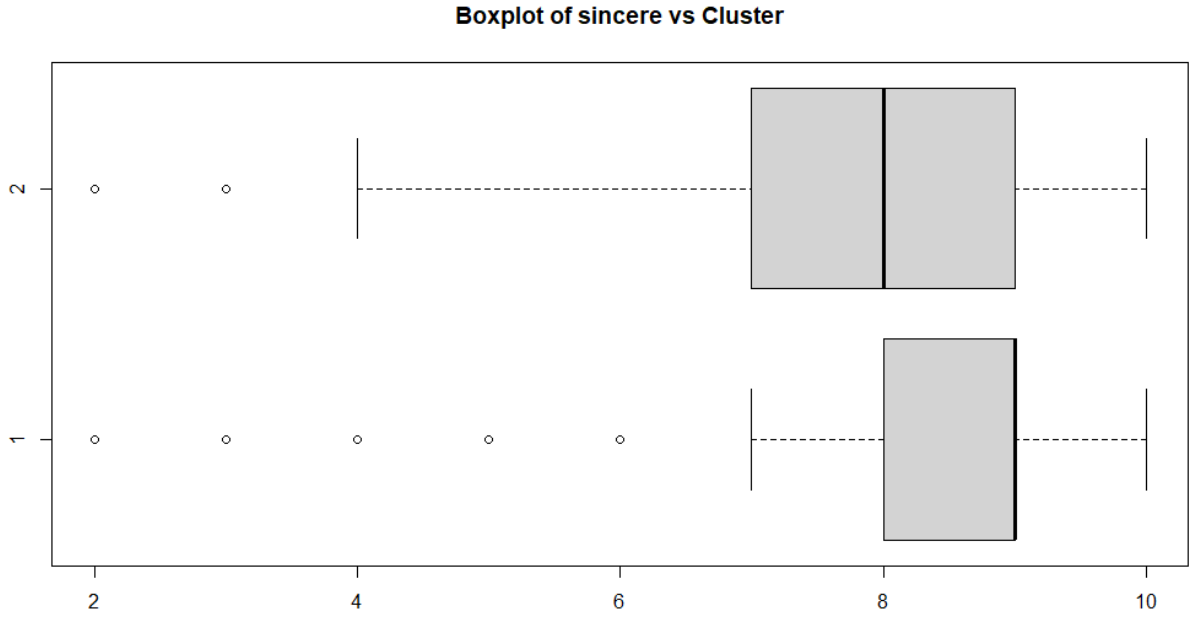
attractive

The next variable we are going to analyze is the relationship between the cluster and variable attractive, which represents the rating about attractiveness made by themselves.
We can observe that the median is the same, but in cluster 1 people tend to evaluate themselves a little bit better, meanwhile on the second cluster the ratings are more distributed.
The difference between the means is 0.814. It's almost 1 point better ranked.
We could conclude that the individuals in cluster 1 are more self confident in terms of attractiveness.



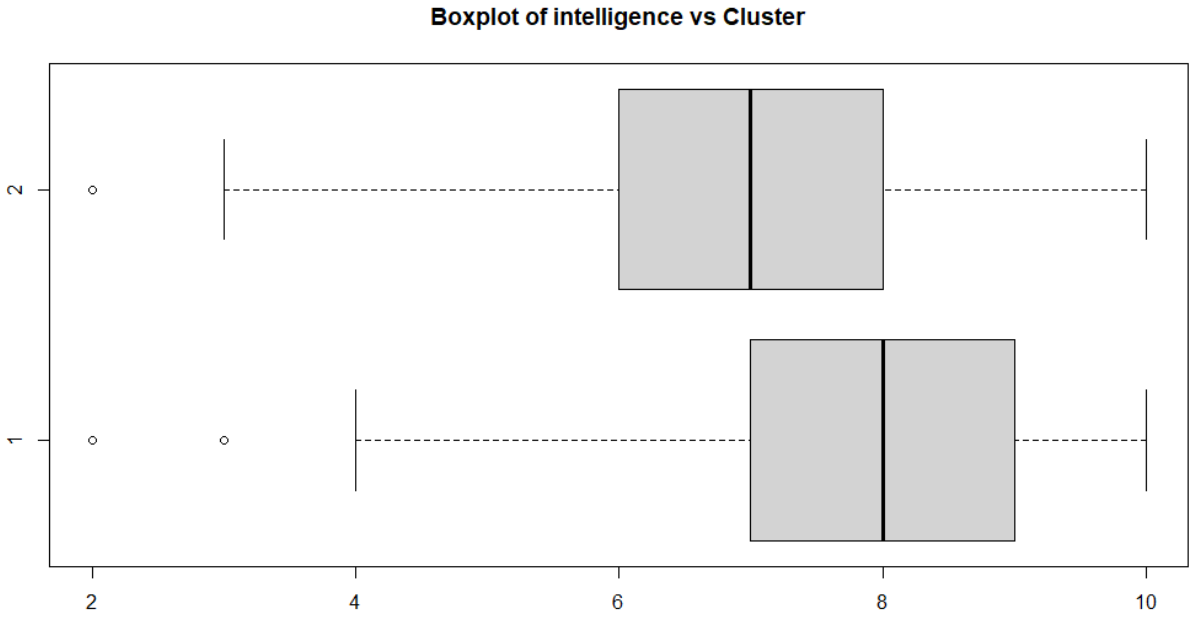
sincere

The next variable we are going to analyze is the relationship between the cluster and variable sincere, which represents the rating about sincerity made by themselves.
We can observe that in cluster 1 people tend to evaluate themselves a little bit better, meanwhile on the second cluster the ratings are more distributed, but still with really good means, above an 8 out of 10.
The difference between the means is 0.251. It's almost nothing.



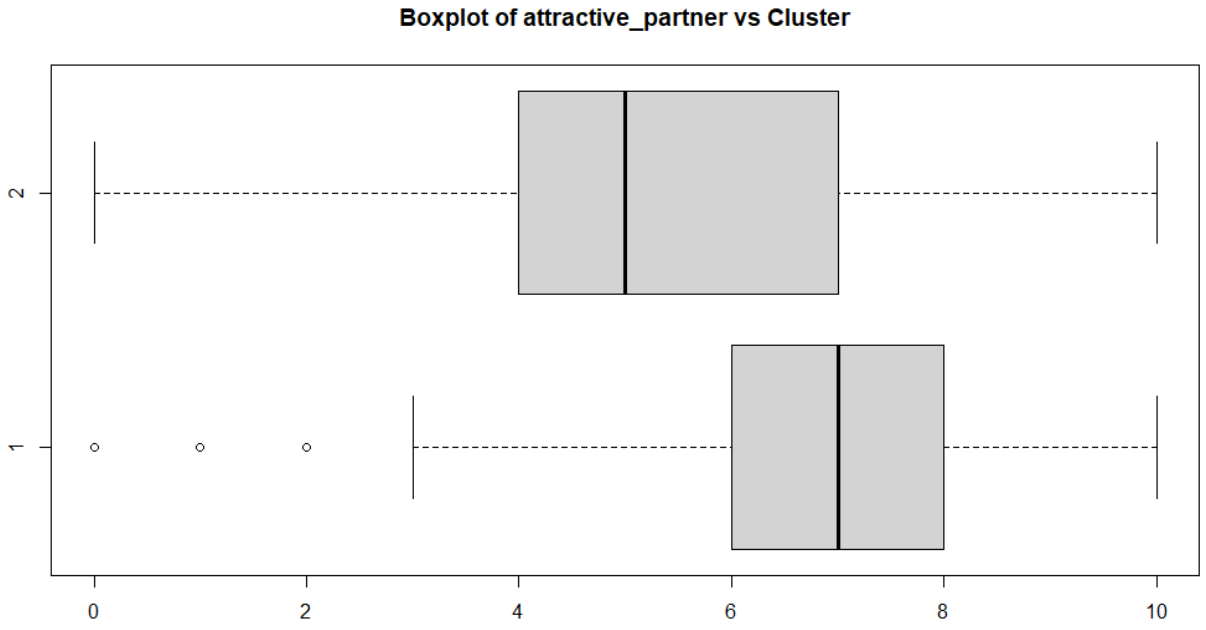
intellgence

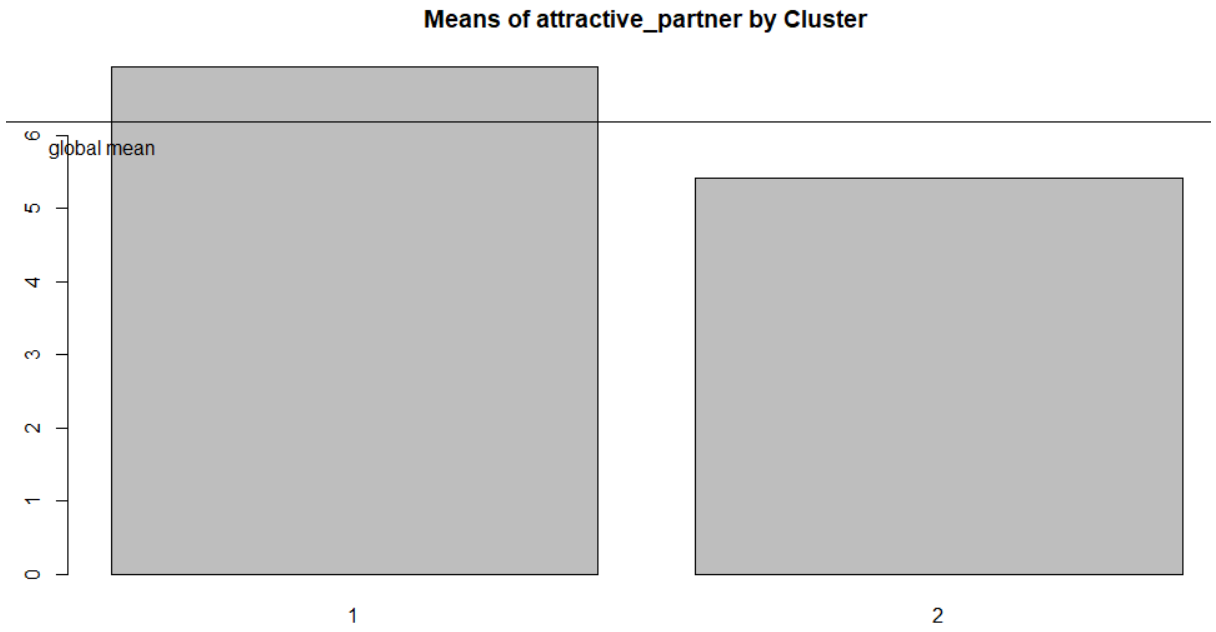
The next variable we are going to analyze is the relationship between the cluster and variable intelligence, which represents the rating about intelligence made by themselves.
We can observe that the individuals in the first cluster are more likely to evaluate themselves better, between 7 to 9. The individuals in the second cluster have ratings a little bit more distributed, with the big amount between 6 to 8.
The difference between the means is 0.94. It's almost 1 point better evaluated.
We could conclude that the individuals in cluster 1 are more self confident in terms of intelligence.



attractive_partner

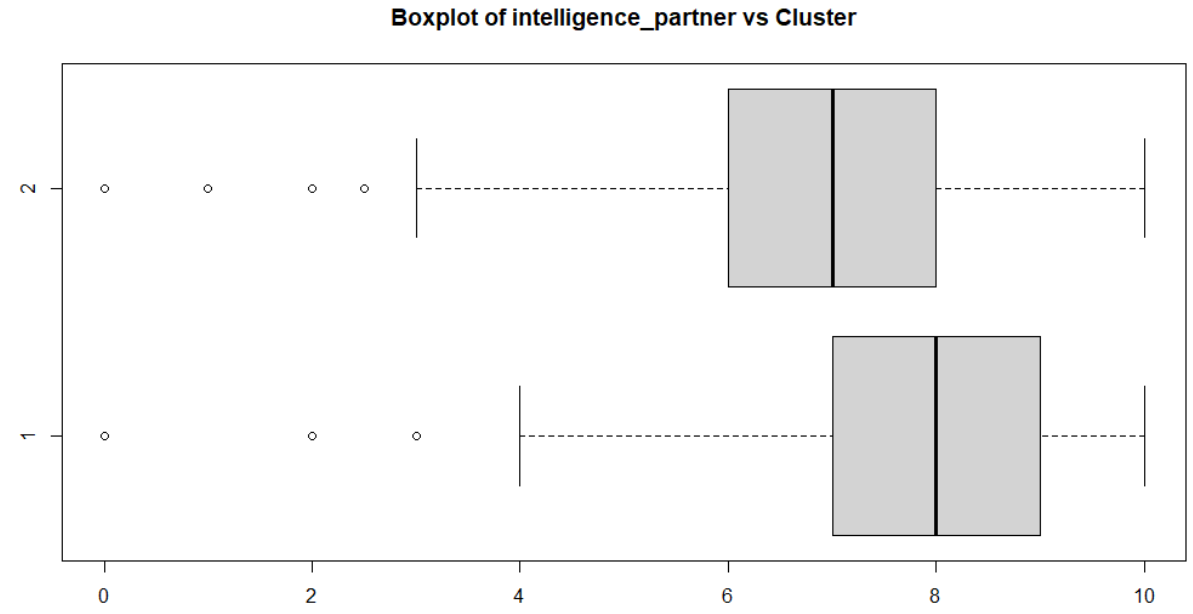
The next variable we are going to analyze is the relationship between the cluster and variable attractive_partner, which represents the rating about attractiveness made by the individual about its partner.
We can observe that the individuals in the first cluster are more likely to rate their partner better, between 6 to 8. The individuals in the second cluster rate more distributed, with the big amount between 4 to 7. We can observe too that there are 2 points of difference in the median.
The difference between the means is 1.527. It's a huge difference of means.
We could conclude that the individuals in cluster 2 are stricter in terms of attractiveness.

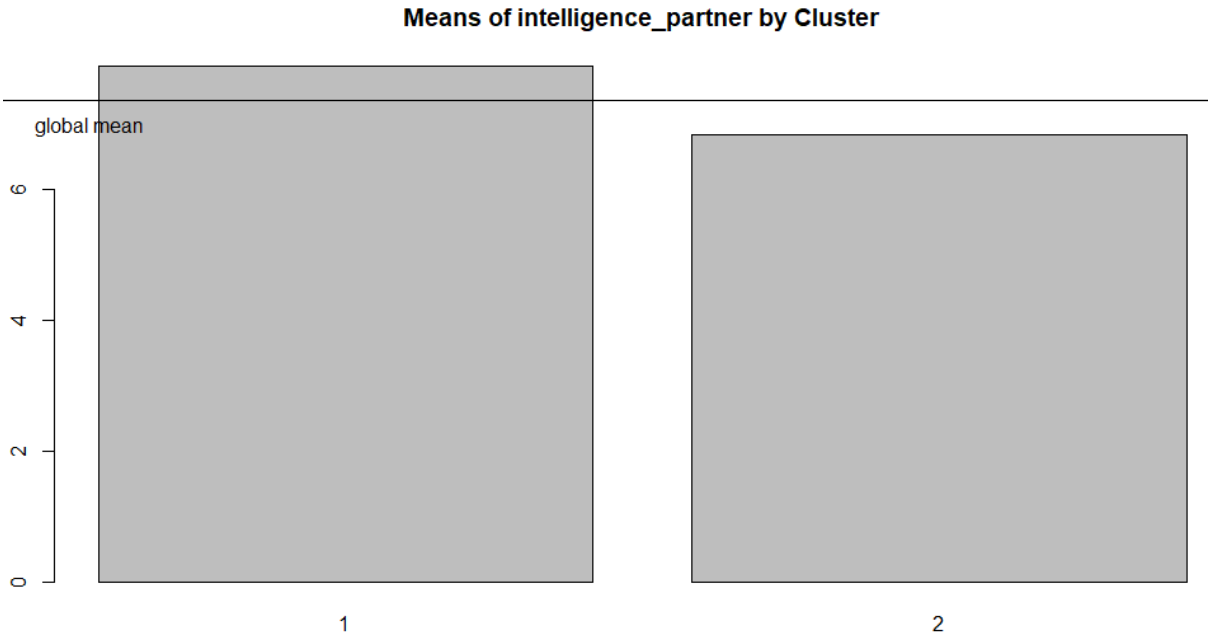




intelligence_partner

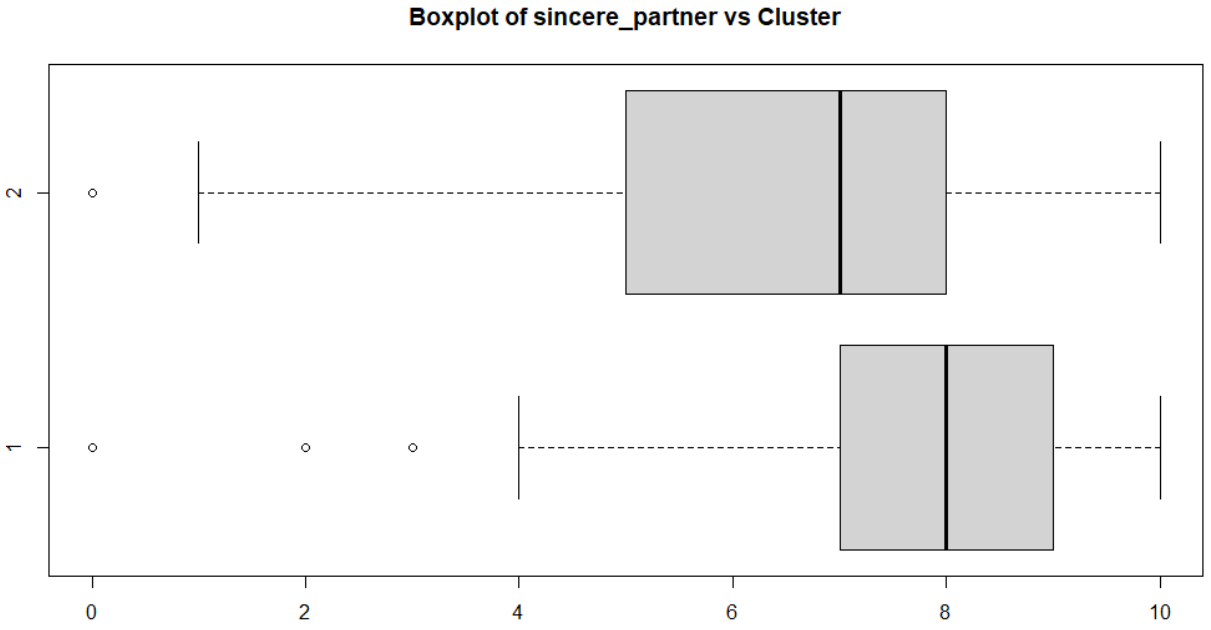
The next variable we are going to analyze is the relationship between the cluster and variable intelligence_partner, which represents the rating about intelligence made by the individual about its partner.
We can observe that the individuals in the first cluster are more likely to rate their partner better, between 7 to 9. The individuals in the second cluster rate more distributed, with the big amount between 6 to 8.
The difference between the means is 1.05. It's a huge difference of means.
We could conclude that the individuals in cluster 2 are stricter in terms of intelligence.





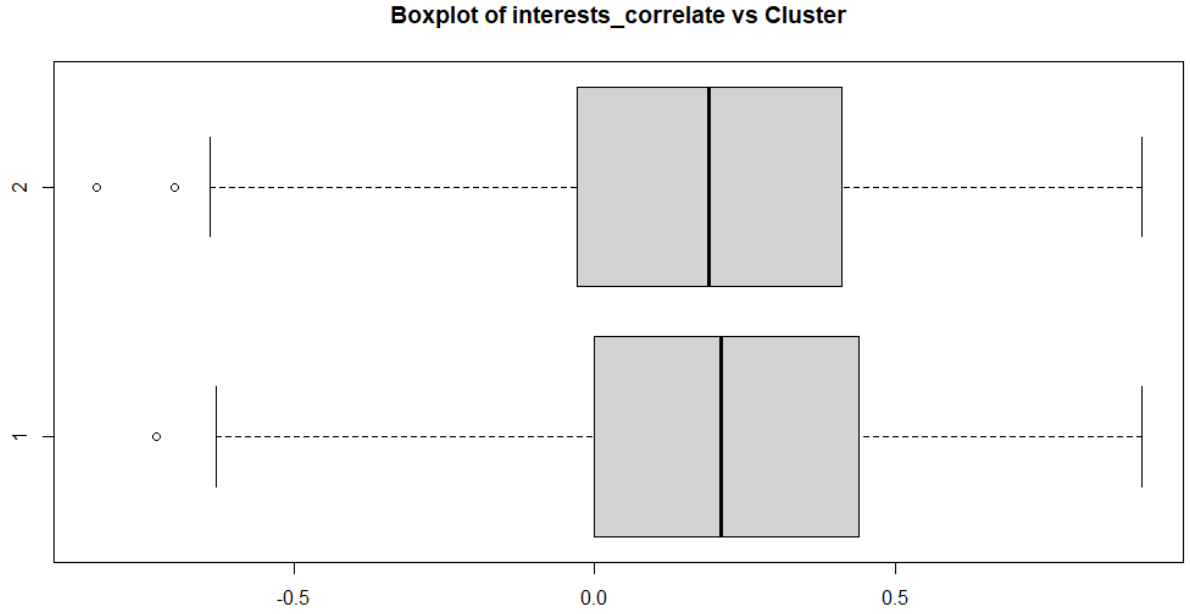
sincere_partner

The next variable we are going to analyze is the relationship between the cluster and variable sincere_partner, which represents the rating about sincerity made by the individual about its partner.
We can observe that the individuals in the first cluster are more likely to rate their partner better, between 7 to 9. The individuals in the second cluster rate more distributed, with the big amount between 5 to 8.
The difference between the means is 1.273. It's a huge difference of means.
We could conclude that the individuals in cluster 2 are stricter in terms of sincerity.



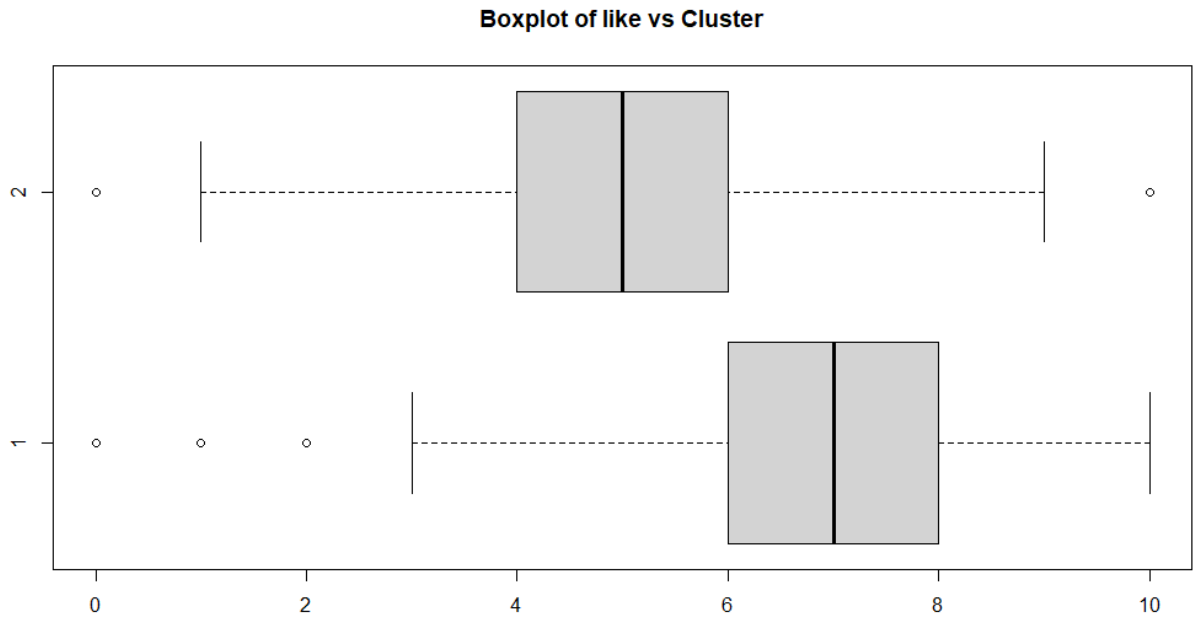
interests_correlate

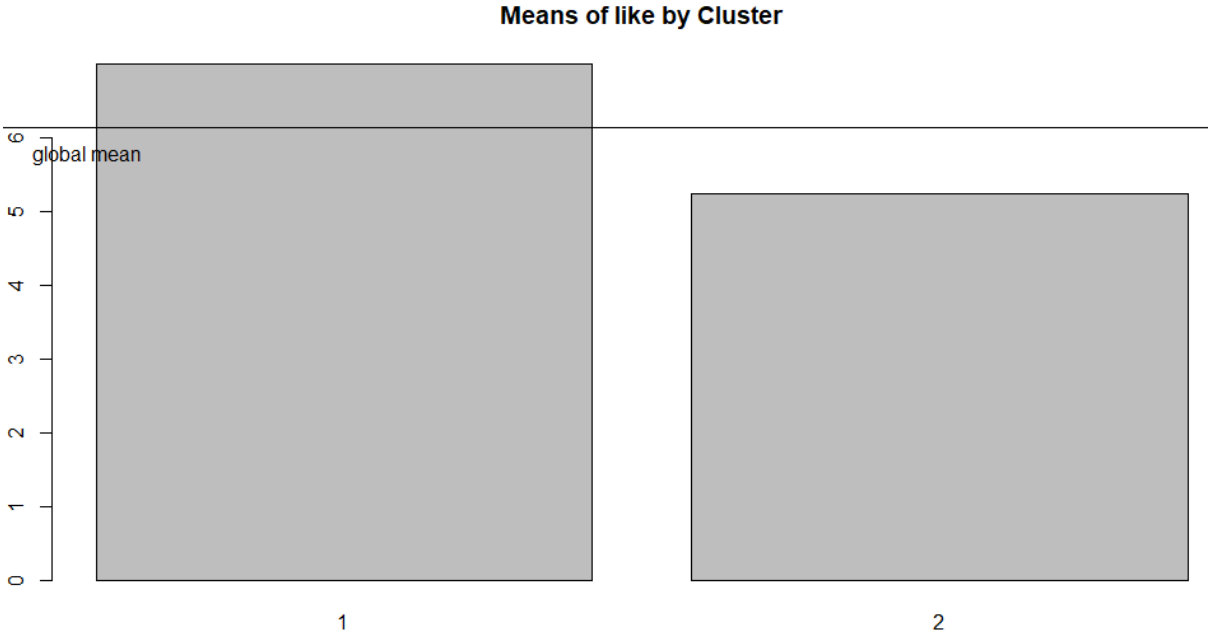
The next variable we are going to analyze is the relationship between the cluster and variable interests_correlate, which represents the correlation between participant's and partner's rating of interests.
We can observe that the two clusters are practically the same.
We could conclude that this variable was not relevant while doing the clustering.



like

The next variable we are going to analyze is the relationship between the cluster and variable like, which represents the rating about how much the participants like their partners.
We can observe that the individuals in the first cluster are more likely to rate their partner better, between 6 to 8. The individuals in the second cluster rate more distributed, with the big amount between 4 to 6.
We can observe a difference of 2 points in the median.
The difference between the means is 1.725. It's almost 2 points of difference.
We could conclude that the individuals in cluster 1 like their partner more.





PCA vs Clustering

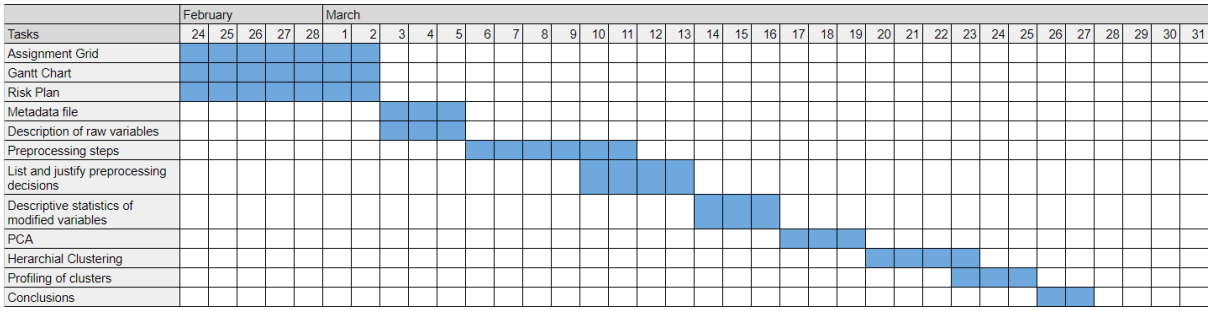
We can see that the results obtained with PCA and Clustering are very similar, in PCA we see that the great majority of individuals accumulate in the centre of the graph, except in cases where there are high ratings. In clustering we can see that there are 2 separable clusters, one cluster belonging to the people who receive no outstanding ratings, and the other one containing the people who receive higher ratings.

Conclusions

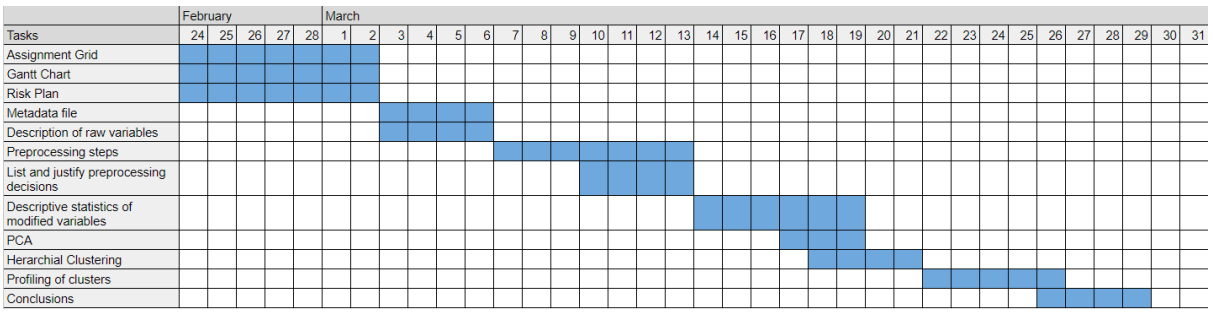
To summarise what we have seen throughout the project, we can say that, although our dataset has not allowed us to provide clear results when clustering or extracting statements, it has shown us that there is a separability in the success that the participants have had, being able to separate them into two groups, one without relevant success and the other that has been very well liked. On the other hand, focusing on the most demanded attributes, we can see that intelligence is the attribute that is given the most importance of all, as opposed to attractiveness, which is not given much importance.

Working plan

Initial Gantt



Final Gantt



Final tasks assignment grid

Members	Maxime Côte	Daniel Muñoz	Pol Pérez	Alejandro Salvat		D2
Assignment Grid		X			1	D3
Gantt Chart	X			**X**	2	D4
Risk Plan			**X**		1	
Metadata file		**X**	X	X	3	
Description of raw variables	**X**	X			2	
Preprocessing steps		X		X	2	
List and justify preprocessing decisions	X		X	**X**	3	
Descriptive statistics of modified variables	X		**X**		2	
PCA			X	**X**	2	
Herarchial Clustering	X	**X**			2	
Profiling of clusters			**X**	X	2	
Conclusions	**X**	X			2	
Coordination	**X**					
	6	6	6	6		